PhD degree in Molecular Medicine (curriculum in Computational Biology)

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Settore disciplinare: MED/04

GENOMICS OF TREATMENT RESPONSE IN ACUTE MYELOID LEUKAEMIA

Margherita Bodini

IEO, Milan

Matricola n. xxxx

Supervisor: Prof. Pier Giuseppe Pelicci

IEO, Milan

Added Supervisor: Dr. Laura Riva

IIT@SEMM, Milan

Anno accademico 2016-2017

Table of contents

List of abbreviations

AML Acute myeloid leukaemia
APL acute promyelocytic leukaemia
bp base pair
CGC cancer gene census
CI confidence interval
CN copy number
CNV copy number variant
CR complete remission
CRi incomplete complete remission
DNA Desossiribonucleaic acid
FAB classification French American British classification
FDR false discovery rate
FN false negative
FP false positive
HF high frequency
HSC hematopoietic stem cell
HSCT haematopoietic stem cell transplantation
indel insertion or deletion
ITD internal tandem duplication
LF low frequency
LSC leukemic stem cell
MDS myelodisplastic syndrome
Mb mega base
MUT mutated
NGS next generation sequencing
NK normal karyotype
PCR polymerase chain reaction
SC stem cell
SNP single nucleotide polymorphism
SNV single nucleotide variant
TCGA the cancer genome atlas
TN true negative
TP true positive
UTR untranslated region
VAF variant allele frequency
WES whole exome sequencing
WGS whole genome sequencing
WHO world health organization
WT wild type

List of Figures

List of Tables

Abstract

Acute Myeloid Leukemia (AML) is a cancer of the myeloid lineage of blood cells characterized by rapid growth of undifferentiated myeloid precursors that accumulate in the bone marrow and suppress normal hematopoiesis. It is the most common adult leukemia with an estimated number of more than 60'000 new cases for the US in 2016. Despite the high rates of complete remissions achieved after treatment (60-80% in young adults), the number of patient that will result cured after induction and consolidation therapy is very low (~12%). The molecular basis of relapsing disease is still unclear and the small number of identified predictive factors has small predictive power. To date, chemotherapy induction treatment is similar for all patients and consists in the administration of mainly three drugs (fludarabine, cytarabine, and idarubicin). Prediction markers for the outcome of chemotherapy would instead reduce useless treatments and direct research through new possible therapeutic targets that would enhance AML treatment. In three successive studies, Ding et al., Corces-Zimmerman et al. and Krönke et al., described four possible behaviors for relapse patients: the return of the first leukaemia (dominant clone or a subclone), with or without additional evolution, or the emergence of ancestral clones, also in this case, with or without additional mutations. In this thesis, endowed of the NGS technologies advancement, we decided to delineate the possible process of relapse formation in order to be able in the future to predict which patients are more susceptible to relapse. Our experimental plan includes the whole exome analysis of 30 pairs of

primary/relapsed AML samples using NGS to identify relapse-specific mutations, the bioinformatics analysis of the clonal evolution of the disease and the identification of pathways that correlate with the relapsing disease.

The methods for the analysis of NGS data, at present, are still in a refinement phase, especially for the high level analysis (detection of variants and definition of their role in the pathogenesis). We broadly analysed the existing methods for the treatment of NGS data (aligners, mutation callers, CNV callers and methods to reconstruct clonal composition) in order to determine those better fitting to our cohort of patients and purposes: occasionally, we had the possibility to choose the best tool meeting our investigative needs, discovering that other methods were valuable as well, in other cases we verified that more improvements are needed to obtain reliable results.

Our analysis shows that the genomic landscapes of primary and relapse AMLs are similar and in the majority of the patients (76%) some relapse clones were already present in the primary tumour and reappeared after chemotherapy at similar or augmented cellular frequencies. We also identified some functional gene categories (DNA methylation pathway, cohesin complex and chromatin modifiers) more prone to resistance and peculiar genes (e.g. ASXL1, TET2) presenting growing VAFs at relapse. In 4 out of 29 patients (14%) we were able to identify driver mutations in the blood sample of the complete remission at low frequency; we hypothesize that more sophisticated diagnostic tools, based on NGS analysis, would help in driving the treatment to obtain better outcomes for patients.

1. Introduction

In the past two decades, the emergence of new technologies for DNA re-sequencing prepared the ground for a deeper knowledge on the characteristics of tumours and the mechanisms of cancer development. In many cases this allowed to opt for more adequate and efficient treatments and better outcome for patients[1,2]. Despite many advances in the Acute Myeloid Leukaemia (AML) genomic characteristics have been disclosed, the treatment and outcome for the majority of patients has not improved. In this chapter we will describe the AML pathology and its emerging characteristics in order to put into context our study on the origin of relapsing leukaemia.

1.0 The blood and the hematopoietic stem cells

The blood serves all the cells of the body for nutrients and oxygen and is responsible for their immune protection. Blood cells have many distinct functions and characteristic morphologies that arise during hematopoietic cell differentiation. Since their lifespan is quite short, ranging from 4-6 days for the platelets to 110-120 days[3] for the red blood cells, there is a continuous need for production of hematopoietic cells by the bone marrow that results in a turnover of ~1 trillion cells per day (for an healthy man of 70 kg[4]) and takes place trough the maturation of the hematopoietic stem cells.

The hematopoietic stem cells (HSCs) are located in the bone marrow and show the typical stem cell characteristics:

- self-renewal: through the mechanism of asymmetric cell division, HSCs are capable of producing new blood cells without the consumption of the stem cell pool, because one of the two daughter cells will differentiate loosing the self-renewal potential while the other will retain all the characteristics of HSCs;

- dormancy: the majority of the stem cells remain in the G_0 (dormant) state of the cell cycle and only a small fraction participate to the active production of blood cells[5] (cytokines are responsible for the activation signal);

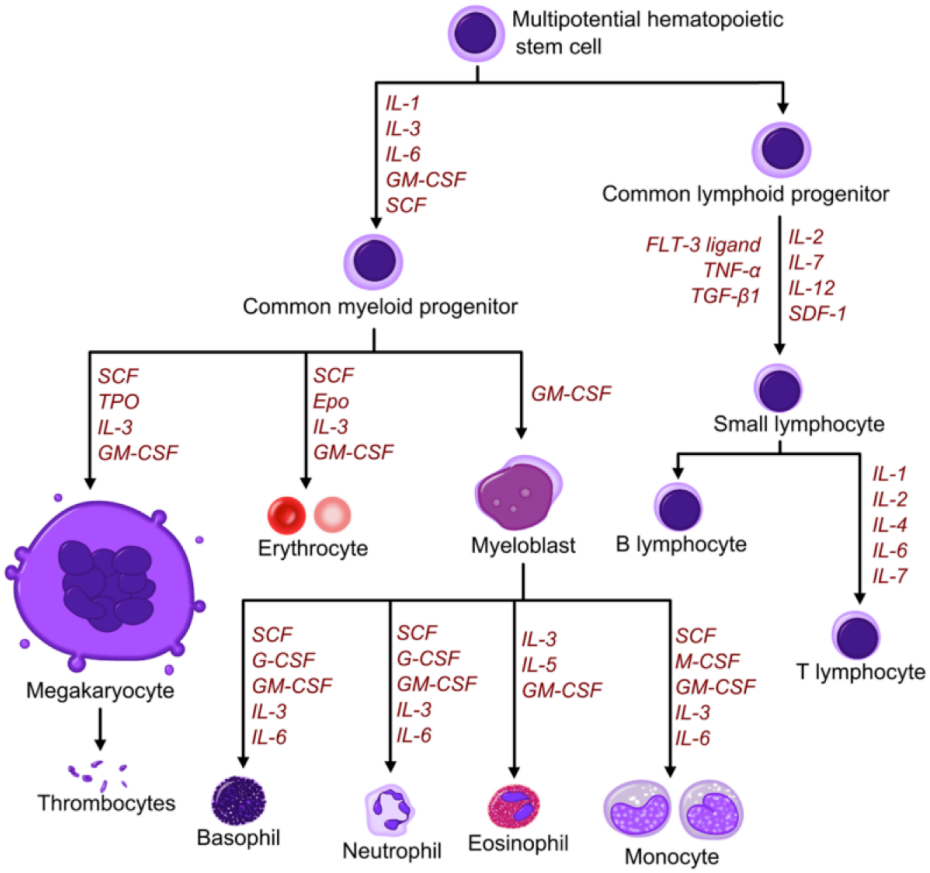
- non specialization: they do not have specific characteristics that allow them to accomplish functions like carrying oxygen molecules or recognize external antigens, but HSCs are capable of giving rise to mature cells.

It is not yet clear whether the differentiation of HSCs is guided by a deterministic or a stochastic mechanism. The former assumption ascribes to the niche the induction of differentiation trough the secretion of specific factors or signalling (e.g. cell-cell interactions or cytokines), whereas the latter envisages complete randomness in the process limiting the role of the niche only to the regulation of which cells are going to progress and which other are dying via apoptosis.[6,7]

The differentiation of HSCs takes place through the expression of a set of genes. The maturation state of a cell is recognizable by the membrane proteins expressed (differentiation markers) and the more it progresses, the more it will be difficult to turning it back to its multipotent primitive state. The cell proliferation is determined by specific growth factors that activate transcription factors via signal transduction pathways. Every depicted branch in Figure 1.1 is associated to specific growth factors (many of which are interleukins) activating the signalling

cascade that leads to a specific mature cell type. Interestingly, some of these growth factors, as erythropoietin and thrombopoietin, find application in the clinics for their capacity to stimulate the production of specific cell types[8]. Successively, transcription factors activate the coordinated expression of groups of proteins; their fundamental role can be gathered by the fact that their mutations are associated to many tumours (MYC is one of the most notorious and studied; for leukaemia, well-known examples are NF-kB and CEBPα).

Figure 1.1: Haematopoiesis in humans. In the graph is depicted the whole lineage of hematopoietic cells along with the factors that promote differentiation at each step.
(Adapted from A. Rad[9])



1.1 Leukaemia

Leukaemia is the cancer of the blood-forming tissue, arising from the abnormal production of blood cells in the bone marrow that divide continuously and eventually substitute normal cells in the blood stream, impeding oxygen distribution to the tissues, immunity functioning and bleeding control.

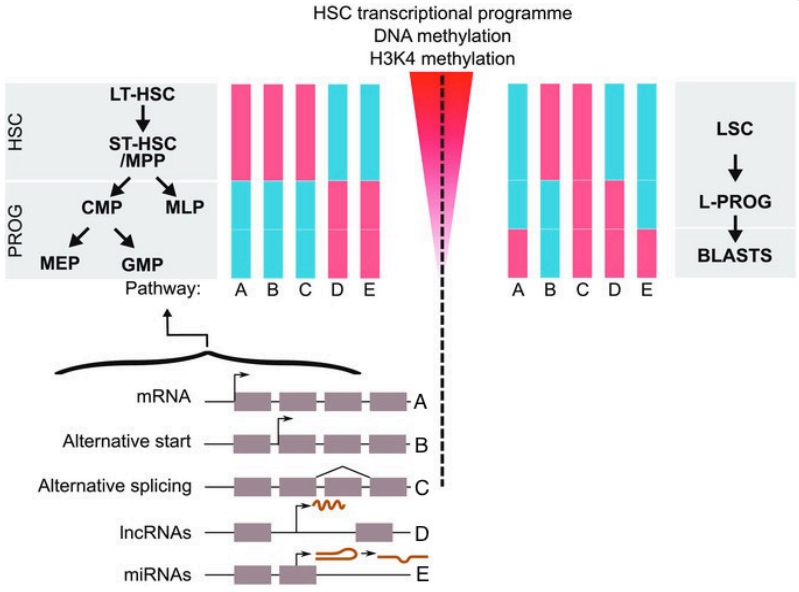
The estimated number of new leukaemia cases in 2016 in the US is 60'140 (3.6% of all new cancer cases, data obtained from NIH[10]); in Europe in 2012 was more than 80'000 cases of which more than 8'000 coming from Italy[11] (as reported by EUCAN). The mortality rate is very high: respectively of 24'000, 54'000, and 6'000 patients per year.

The program of differentiation defined in normal haematopoiesis is significantly modified in leukaemia. In normal haematopoiesis HSC differentiate in all the cellular components of the blood, in leukaemia the stem cells are blocked in their maturation program in a state defined as blasts. The subgroup of the blasts owing all the stemness characteristics is called Leukemic Stem Cells (LSCs) and is thought to serve as a reservoir of leukemic cells acting similarly to SCs for normal tissues. Theoretically, these cells are responsible to repopulate the leukaemia in immunodeficient mice upon transplantation. Additionally, it is believed that their quiescent phenotype in some cases provides them the capacity to survive chemotherapy and relapse. Gene expression defines the phenotypical nature of a cell, and maturation is pursued switching on and off specific groups of genes. As depicted in Figure 1.2, in normal haematopoiesis (left panel) there is a clear

definition of biological pathway activation that leads from stem cells to progenitors; the same happens in leukemic cells although in a drastically different manner.[12] LSCs have simultaneously strong similarities with HSCs and normal progenitors; again, differentiated blasts share many pathway characteristics with normal hematopoietic cells at diverse maturation stages. In leukaemia the organization of modules activation and inactivation is mixed.

Figure 1.2: Activation and inactivation of cellular function pathways in normal HSCs and LSCs. Differentiation from stem cells to progenitors in the bone marrow follows a specific pattern of activation/inactivation of biological processes that leads to a certain transcriptional program. In LSCs that pattern is consistently modified in the maturation from stem cells to leukemic blasts.

(Adapted from Vedi et al., 2016[12])



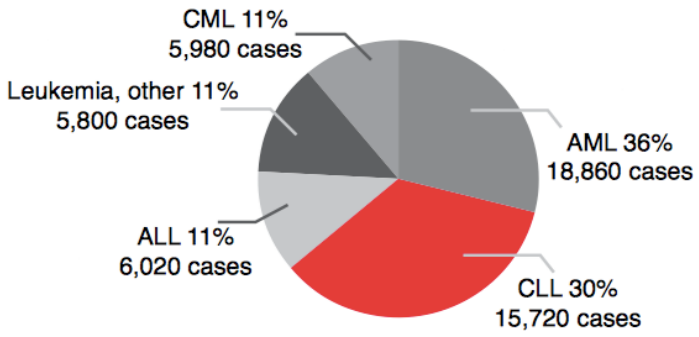
Leukaemias are grouped into 5 major types, based on the cell type that became cancerous (lymphoid or myeloid) and on the progression of the disease, that can be rapid (acute) or slow (chronic):

- Acute Lymphoblastic Leukaemia (ALL),
- Acute Myelogenous Leukaemia (AML),
- Chronic Lymphocytic Leukaemia (CLL),
- Chronic Myelogenous Leukaemia (CML),

- Other leukaemias.

Relative percentages of these 5 groups, as estimated for USA in 2014, are shown in Figure 1.3.

Figure 1.3: Proportion of new leukaemia cases, divided per type, in USA in 2014: an estimation done by the American Cancer Society.
(Adapted from Cancer Facts and Figures, 2014)



In this study we are focusing our attention on the AML subgroup, which is the most frequent among adults; therefore, we are describing its characteristics more in detail. AML involves the myeloid lineage of hematopoietic cells and is characterized by the rapid growth of white blood cells in the bone marrow. Even though great advances have been made in the genetic characterization of this tumour, the disease entities are defined predominantly on cytogenetic and molecular markers. Two major classifications have been proposed and are used today to classify AML patients, the former is the French-American-British (FAB) classification firstly delineated in 1976, that groups AML subtypes based on cell-type and maturation of the cell of origin[13] (Table 1.1); the latter was produced by the World Health Organization (WHO) in 2008[14] and revised in 2016[15] and aims at a major relevance to the clinics and treatment (Table 1.2).

Table 1.1: French-American-British categorization of AMLs with cell-type and maturation of the cell that originated the disease and typical cytogenetic traits for each group.

FAB subtype	Name	Cytogenetics
M0	Undifferentiated AML	
M1	AML with minimal maturation	
M2	AML with maturation	t(8;21)(q22;q22), t(6;9)
M3	APL	t(15;17)
M4	Acute myelomonocytyc leukaemia	inv(16)(p13q22), del(16q)
M4 eos	Acute myelomonocytic leukaemia with eosinophilia	inv(16), t(16;16)
M5	Acute monocytic leukaemia	del(11q), t(9;11), t(11;19)
M6	Acute erythroid leukaemia	
M7	Acute megakaryoblastic leukaemia	t(1;22)

Table 1.2: WHO schematic representation of subclasses of AML (2016).

Acute myeloid leukaemia (AML) and related neoplasms
AML with recurrent genetic abnormalities
AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1
AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22);CBFB-MYH11
APL with PML-RARA
AML with t(9;11)(p21.3;q23.3);MLLT3-KMT2A
AML with t(6;9)(p23;q34.1);DEK-NUP214
AML with inv(3)(q21.3q26.2) or t(3;3)(q21.3;q26.2); GATA2, MECOM
AML (megakaryoblastic) with t(1;22)(p13.3;q13.3);RBM15-MKL1
Provisional entity: AML with BCR-ABL1
AML with mutated NPM1
AML with biallelic mutations of CEBPA
Provisional entity: AML with mutated RUNX1
AML with myelodysplasia-related changes
Therapy-related myeloid neoplasms
AML, NOS
AML with minimal differentiation
AML without maturation
AML with maturation
Acute myelomonocytic leukaemia
Acute monoblastic/monocytic leukaemia
Pure erythroid leukaemia
Acute megakaryoblastic leukaemia
Acute basophilic leukaemia
Acute panmyelosis with myelofibrosis
Myeloid sarcoma
Myeloid proliferations related to Down syndrome
Transient abnormal myelopoiesis (TAM)
Myeloid leukaemia associated with Down syndrome

Cytogenetic is used for classification but serves also as a prognostic factor. In fact, cytogenetic abnormalities have been associated to different relapse risks and overall survival. Grimwade et al.[16] defined 3 risk classes: favourable, intermediate and adverse risk, based on the AML cytogenetics. Also genetic mutations have been associated with AML risk and principal markers are described in the next paragraphs.

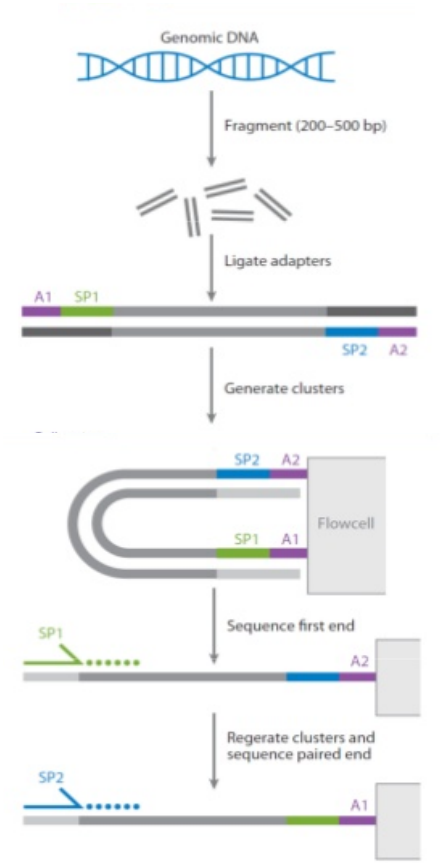
1.2 Next Generation Sequencing

The advent of Next Generation Sequencing (NGS) allowed researchers to sequence genomes of already assembled organisms in a more rapid and cheaper manner than before. It is often referred as high throughput sequencing because its approach endeavours at the repeated sequencing of the same regions in order to strengthen the power of base calls. It can be used only for already assembled or very little genomes because it produces short reads that need to be aligned to a reference in order to map their position. In particular, the technology we used to sequence the human genome is the Illumina platform (Solexa), the bases of which are schematized in Figure 1.4. The genomic DNA is fragmented in small pieces (generally, it is sonicated and the fragments have a maximum size of 500 bp); after fragmentation, specific adapters are ligated to each end of the DNA fragments. Through the adapters, each fragment is fixed on the sequencing flow cell and clusters of fragments are formed by PCR. The sequencing can be

performed twice, starting from both ends of each fragment for a specified number of nucleotides. The sequencing output consists of millions of sequencing reads of a given length (75, 100 or 150). The sequencing technique uses reversible terminator nucleotides: at each base the four nucleotides labelled with different fluorochromes can be incorporated, after wash out of the not incorporated nucleotides, the luminescence is registered. By subsequent rounds of sequencing, for every cluster present in the flow cell, forward and reverse sequences are produced. The sequencing can be pursued throughout all the genome (whole genome sequencing, WGS) or on targeted regions. Whole exome sequencing (WES), for example, targets all the known exons in the genome thanks to a specific capture enrichment step that uses oligonucleotides specifically designed to select only the portions of the genome of interest.

Figure 1.4: Illumina sequencing. In the figure are depicted the steps needed to perform Next Generation Sequencing through the Illumina platform. The genomic DNA is fragmented and adaptors are ligated to each end. The fragments are fixed to the flow cell trough the adapters and the clusters are generated by successive PCR cycles. Afterwards, both end of the fragments are sequenced.

(Adapted from Illumina)



The output of the machine consists in raw files containing all the reads information. The mutational analysis follows successive steps, starting with the mapping of all those reads to the reference genome through alignment algorithms. The reference genome derives from the sequencing of the genome of a single individual and, although many enhancements have been made, reads coming from another healthy individual aligned to the reference genome will, naturally, differ from it in many positions. These positions are called Single Nucleotide Polymorphisms (SNPs): they are the portions of the genome that characterize the phenotypical differences between individuals and, aside for some cases in which they are predisposing for some diseases, they do not have a

malignant potential. On the contrary, it happens in cancer that a genomic position differs somatically in the tumour cells from the correspondent germline position of the same individual: these positions are called single nucleotide variants (SNVs) and can be responsible for malignant phenotypes. The same applies for small insertions or deletions (indels) meaning that one or more bases can be gained or lost somatically. In the cancer context, the frequency at which this allelic differences are found is fundamental because it reflects the portion of cells carrying that variant in the tumour. The Variant allele frequency (VAF) is calculated as the number of reads carrying the alternative variants among all the reads spanning that genomic position.

1.3 AML genomic landscapes

As well as in many other cancer types, mutations implicated in the pathophysiology of AML may cause the activation of a proto-oncogene, the inactivation of a tumour suppressor gene or can alter the transcription of a gene through the mutation of the transcription factors binding sites. Thanks to the rapid evolution of technology in the last twenty-five years the mutational status and the levels of expression of many genes have been linked to acute myeloid leukaemia, however their exact coordination in the development of the disease still needs to be uncovered. We are, here, describing what is known up to date about the AML genome, retracing the milestone discoveries of almost thirty years of cancer research.

1.3.1 Established alterations in AML

Cytogenetic alterations has been used for a long time as the major distinctive factor between AML subgroups, yet around half of AMLs have normal karyotype.[17] In this paragraph we describe the major genetic alterations that have been discovered in AML through molecular biology before the advent of NGS. Since this study focuses on SNVs and indels landscapes, we are not explaining relevant translocations like AML1-ETO and PML-RARa, which are distinctive for AML subgroups.

1.3.1.1 FLT3

At the beginning of the 90's a novel receptor tyrosine kinase was discovered to be specific of murine haematopoietic stem cells with enriched stem cell activity[18]. Further studies showed that the Fms-Related Tyrosine Kinase 3 (FLT3) is a growth factor receptor that promotes autologous proliferation, and it is quite commonly found constitutively activated in leukaemia patients. The most common mutations of FLT3 consist in an internal tandem duplication (ITD) that is found in 20-25% of the patients, and in a point mutation in the tyrosine kinase domain, that occurs in 7.7% of the patients. The mutated allele is not always expressed and patients with a high mutant *versus* wild type ratio display a shorter overall and disease free survival[19].

1.3.1.2 NPM1

Nucleophosmin (NPM1) is a nucleolar phosphoprotein that shuttles between the nucleus and the cytoplasm. In 2002 NPM1 was described as a crucial regulator of the tumour suppressor p53[20]. Variants in NPM1 genes are generally insertions of 4 bp in the 12th exon, causing a frame-shift that alter the C-terminal of the protein. The insertions most frequently observed, in descending prevalence order, are: variant B (960insCATG), variant C (960insCGTG) and variant D (960insCCTG). NPM1 variants are specifically observed in AML (35.2% of cases) and not in other neoplasms (both haematopoietic or extrahaematopoietic)[21] and they confer distinct features to the mutated patients from a genetic, clinical, pathologic, immunophenotypic and cytogenetic point of view[22] (Table 1.3). In particular, they are associated with a specific AML subgroup and display a gene-specific homeobox expression signature.[23] NPM1 mutations can occur together with FLT3 mutations, in this case the prognosis for the patient is better than for FLT3 mutation alone.

Table 1.3: Characteristic features of NPM-mutated AML as reported by Falini et al.[21]

Genetic features
NPM1 mutation is specific for AML, mostly "de novo"
Usually all leukemic cells carry the NPM1 mutation
Mutually exclusive with other "AML with recurrent genetic abnormalities"
NPM1 mutation is stable (consistently retained at relapse)
NPM1 mutation usually precedes other associated mutations (e.g. FLT3-ITD)
Unique GEP signature (↓ CD34 gene; ↑ HOX genes)
Distinct microRNA profile
Clinical, pathologic, immunophenotypic, and cytogenetic features
Common in adult AML (~ 30% of cases), less frequent in children (6.5%-8.4%)†
Higher incidence in female
Close association with normal karyotype (~ 85% of cases)
~ 15% of cases carry chromosome aberrations, especially +8, del9(q), +4
Wide morphologic spectrum (more often M4 and M5)
Frequent multilineage involvement
Negativity for CD34 (90%-95% of cases)
Good response to induction therapy
Relatively good prognosis (in the absence of FLT3-ITD)

1.3.1.3 RAS family

The Ras family is a group of proteins with GTPase activity that play a vital role contributing in regulation of cell proliferation. In fact, several members of the family are mutated in many cancer types and they have been associated also to leukaemia in the late eighties as an infrequent mutation (~25% of the patients).[24,25] Three notorious proto-oncogenes, that express the $p21^{RAS}$ protein, are part of this family: HRAS, KRAS and NRAS.[26] Ras-family mutations generally affect codons 12, 13 and 61 of the gene transcript causing the amino acid change of a glycine into a bigger residue (valine, aspartate, cysteine, serine).[27] KRAS mutated patients respond better to high cytarabine doses than not mutated patients[28]; its pathway or its regulating components are often associated with MLL rearrangements[29] and are probably the result of genomic instability that causes damages in the most fragile portions of the genome.

1.3.1.4 CEBPA

CEBPA is a transcription factor that coordinates proliferation and differentiation in myeloid progenitors, fine-tuning the activity of cyclin-dependent kinases. CEBPA variants often result in the disruption of the leucine zipper domain or in the premature protein termination and can be generated both by indels or point mutations affecting protein translation.[30,31] Also a familial mutation has been reported for CEBPA, consisting in the deletion of the 212C.[32] Variants in CEBPA occur in 16% of the patients[30] and are associated with a good prognosis.

1.3.1.5 RB1

RB1 gene has a role in regulation of proliferation and differentiation, acting directly on the cell cycle and interacting with p53, MDM2 and the polycombs. Its role in cancer was primarily assessed for the retinoblastoma neoplasms, but mutations of RB1 are found also in AML. However, in contrast with the point mutations associated to the retinoblastoma phenotype, in leukaemia the RB1 gene is often target of gross rearrangements and has been associated to poor prognosis.[33]

1.3.1.6 TP53

TP53 is probably the most notorious tumour suppressor gene. Its role is fundamental in senescence and apoptotic responses. Cells with dysfunctional p53 can be subjected to wrong rearrangements of short chromosomes that lead to

the production of circular chromosomes and eventually to the wrong redistribution of genomic DNA after cell duplication. In fact, in leukaemia p53 mutations are often associated to complex karyotypes.[33] As long as the majority of tumour suppressors, mutations in TP53 do not need to be positioned exactly at one specific nucleotide of the gene: generally they are sparse and located between exon 5 and 8 (especially on the last). Mutated TP53 usually confer a worse prognosis in AML[34].

1.3.2 The genomic era

In 2008 the DNA from a patient affected by AML has been the first cancer genome to be completely sequenced. For this study, the authors chose one patient presenting an AML belonging to the M1 subtype, which is the most common, representing approximately the 20% of all cases. The authors selected this particular subtype, thinking that its genome would be easier to analyse, as first attempt, because it is not associated with common genetic abnormalities. The authors identified the mutations by comparing the DNA coming from the tumour sample to its correspondent skin sample, used as normal counterpart, and discarding all the variants hitting non-genic regions to slim down the list. The resulting 11'192 variants were again refined to exclude introns, not-translated and synonymous variants (i.e. variants that do not results in changes of the amino acidic composition of a protein) to obtain a final list of 181 possible mutations. The majority of them resulted to be false positives. Indeed, the authors were able

to validate by PCR and Sanger sequencing only a list of 8 single nucleotide mutations and 2 small indels. These variants were present both in the tumour and the relapse sample (not sequenced, but tested for these variants) and were nearly absent in the skin sample, as expected.

The two indels were common mutations in leukaemia, thus found in the FLT3 and the NPM1 genes. On the contrary, the eight Single Nucleotide Variants (SNVs) had not been previously identified in any leukemic genome and were absent in a cohort of 187 leukemic patients that were analysed for these mutations by PCR. However, the authors speculated that the inability to retrieve these mutations in other AML patients was mainly due to the small dimension of the test cohort and speculated on the possibility for PTPRT, CDH24, PCLKC and SLC15A1 to be implicated in the pathogenesis of AML, on the basis of the recurrence of mutations in these genes in cancers other than leukaemia. Also the other 4 genes, KNDC1, GPR123, EBI2 and GRINL1B, were interesting for their potential function. Notably, the skin sample, sequenced as normal reference, was found contaminated by leukemic cells, anticipating the important issue of the choice of which tissue to use as normal sample for genome and exome analysis of blood tumours[36].

After this seminal paper, many others studied the genomic landscape of AMLs and uncovered the relationships between genes and this pathology. The most important players newly identified are IDH1, IDH2,[37] TET2,[38] DNMT3A,[39] EZH2[40] and UTX[41]. Recently, also the spliceosome machinery have been associated to leukemogenesis[42]: mutations of spliceosomal genes, which fall in the class of

tumour suppressors, may cause defective splicing, resulting in accumulation and, as a consequence, unbalanced ratios of isoforms of crucial genes in the cells (e.g. RUNX1) that lead to leukaemia[43].

There are many groups worldwide aiming to characterize AML at a genomic, epigenomic and transcriptomic level and a joined effort has been made to build public databases of samples and clinical information from haematological patients. An example is the GIMEMA foundation in Italy, which collects samples and information from several haematological units all around the country; started with the aim to develop scientific research on haematological malignancies, now serves as a reservoir of samples with associated clinical information.

The first genomic and epigenomic study on a relatively big number of AML patients was published by The Cancer Genome Atlas (TCGA), that analysed the DNA, either genomic or exomic, from 200 *de novo* AML patients in order to delineate the commonalities and peculiarities of this pathology[44]. This dataset still represents the largest public collection of AML DNA data and many successive publications performed further analysis on this same dataset.

Compared to many other cancer types, AML results to be one of the less mutating tumours, with a median mutation frequency across patients of 0.37 *per* Megabase, even if there is a high inner variability among individuals, with the range of mutations that spans two orders of magnitude from 0.01 to 10 *per* Megabase[45]. Lawrence et al. in their 2013 Nature paper[45] were able to correlate the genomic frequency of mutations both with the replication timing and the transcriptional activity across the genome, the overall small number of mutations

identified in AML patients may very likely be due to a slow cycling activity of the tumour cells of origin. Therefore, considering the natural error rate of the replication machinery and the replications time of such cells, a diminished mutation frequency can be imputed to a lower probability of accumulating mutations. However, the reason for the high variability of mutation rates across patients in AML remains still an open issue.

Despite the cohort of patients was quite small in order to make good association studies, the authors observed a significant low number of mutations in patients presenting the PML-RARa and the MLL-x translocations compared to other leukaemia types, suggesting that these abnormalities do not need many cooperating events during the leukemogenic process. On the contrary, AMLs with RUNX1-RUNX1T1 fusions, with TP53 mutation and the AML associated to unfavourable risk groups present a significantly augmented number of mutations.

Tier1 mutations are those causing changes in the amino acid coding regions of annotated exons, consensus splice-site regions, and RNA genes (including microRNAs).[44] In the study of the TCGA mentioned above, AML presented from 0 to 51 Tier1 mutations *per* patient with an average of 13 mutations. The analysis was carried on 200 AML patients for a total of 2315 somatic variants, 1528 (66%) of which were missense and 270 small insertion and deletions, which caused frameshifts in 192 (71%) cases.

Recurrent genes are defined as the genes mutated somatically in at least two samples: 260 genes presented these characteristics and 154 of them were recurrently mutated missense. Among them, the authors scored genes already

established to have a role in the pathogenesis of AML from a very long time (DNMT3A, FLT3, NPM1, IDH1, IDH2 and CEBPA) and from recent years (U2AF1, EZH2, SMC1A and SMC3).

1.3.2.1 Driver and passenger mutations

It is straightforward that distinct mutations may have disparate relevance for cancer formation depending on the base of the codon they affect: they can have no effect (synonymous mutations), cause the change in amino acid sequence (non synonymous mutations) or determine the complete termination of the amino acid chain (nonsense mutations). Furthermore, the gene that has been targeted by the mutation can be either expressed or not in the cell type expanding in the tumour and the gene can have an essential or irrelevant role for the maintenance of regular cellular functions (many cellular pathways enclose redundancies). The mutations that are responsible for the tumour phenotype are called "driver", while other mutations that are carried in tumour cells but do not give significant contribution to the altered condition are called "passenger".

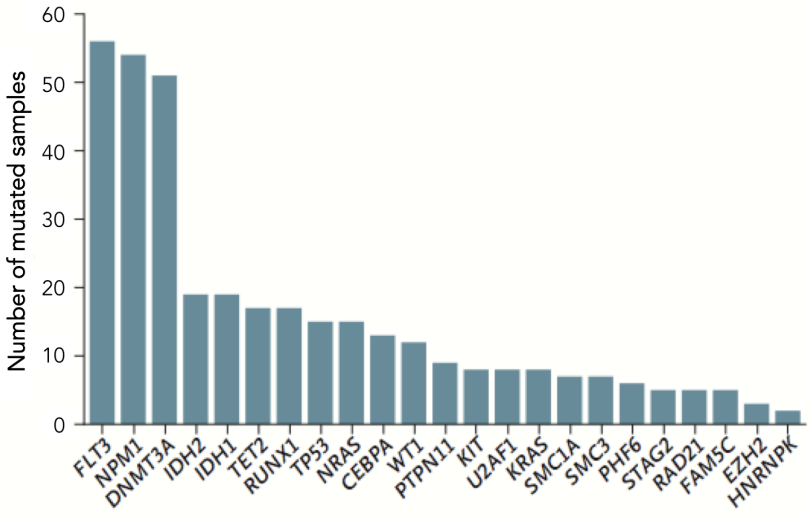
Many tools have been developed to recognize which genes are more likely to have a paramount role in cancer development thus including "driver" mutations (i.e. "driver" genes). Driver genes are thought to have an enriched frequency in the patient population because of their causality. Therefore, algorithms that aim at the identification of driver genes use for their predictions the information about the frequency of the mutation in the patient populations, in some cases corrected for other confounding parameters[45] (e.g. gene length), or in other cases

associated to other quality information such as the typical mutational patterns of an oncogene or a tumour suppressor gene, like DOTS-Finder, a tool developed in our laboratory.[46]

In Figure 1.5 are reported the 23 genes which resulted to have a significant mutation prevalence from the analysis with Mutational Significance in Cancer (MuSiC) tool[47], that aims at the distinction of true causal mutations from passenger events. The majority of them had been previously described as crucial players in AML development because they were molecularly identified in AML patients.

Figure 1.5: Genes significantly prevalent in AML according to MuSiC. In the graphic is reported the number of mutations observed for the 23 genes identified as drivers using the MuSiC tool on the TCGA data on AMLs.

(Adapted from TCGA, 2012[44])



1.3.2.2 Functional categories of genes implicated in AML

The process that drives from mutations to tumour development is due to the impairment of essential cellular functions. Different mutations in different genes can give the same phenotypical outcome if they affect the same pathway. The

investigation of altered pathways may, therefore, uncover the fundamental functions that are disregulated in the specific tumour (i.e. AML); consequently, genes that would be catalogued as passenger from an approach based uniquely on frequencies, can unveil their causal role when investigated in their pathway context. HotNet[48] is the implementation of an algorithm designed to identify de novo subnetworks implicated in tumour development, given a list of mutations detected in a cohort of patients. Starting from an interaction network, HotNet firstly defines gene neighbourhoods in a diffusion perspective and, subsequently, tests the false discovery rate of the identified subnetworks in order to extract the more reliable. Using this tool, the TCGA[44] described that 99% of AML patients carried at least one mutation belonging to one of the 9 clusters identified (Figure 1.6):

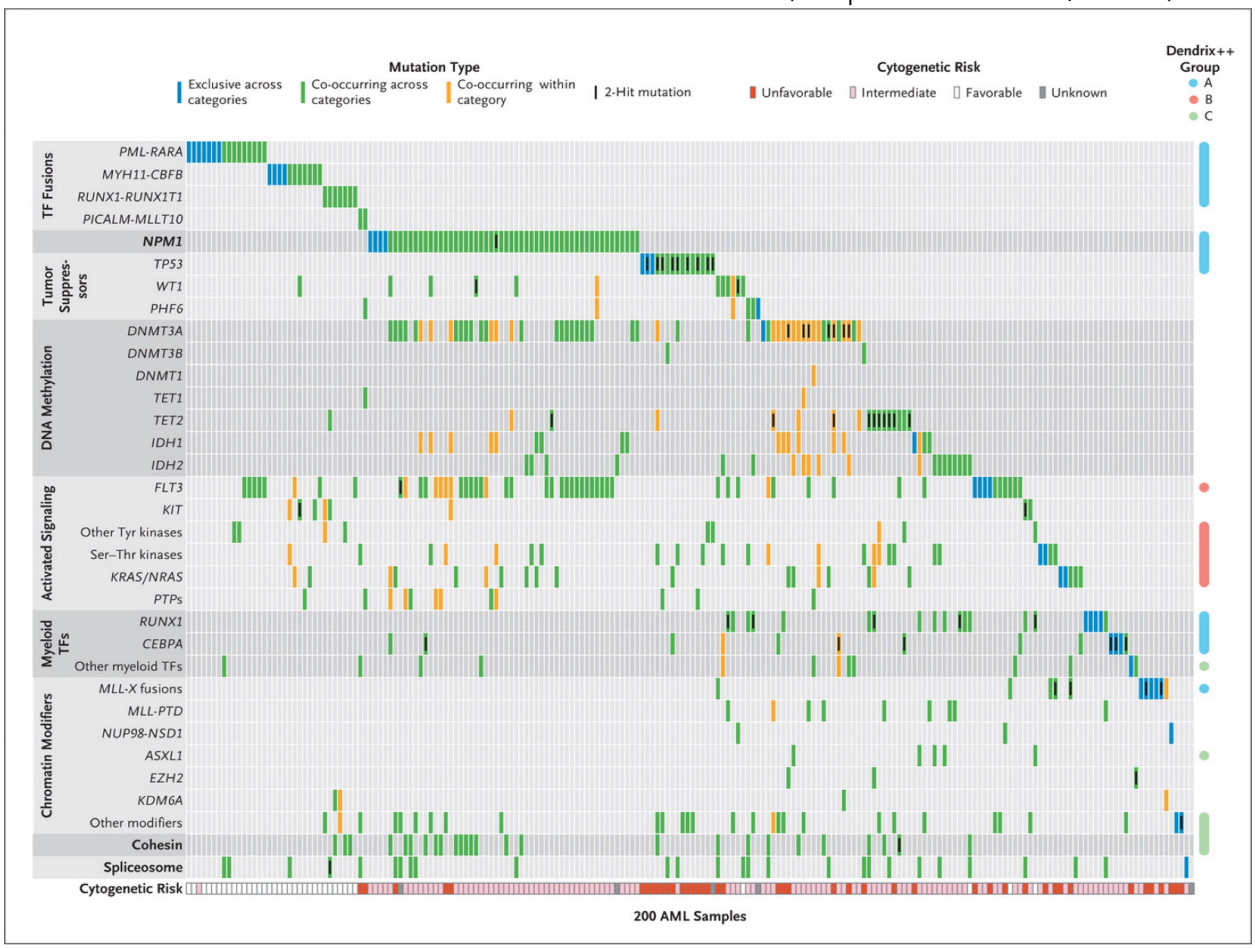
- transcription-factor fusions,
- the gene encoding for nucleophosmin (NPM1),
- tumour suppressor genes,
- DNA-methylation–related genes,
- activated signalling genes,
- chromatin-modifying genes,
- myeloid transcription-factor genes,
- cohesin-complex genes,
- spliceosome-complex genes.

Furthermore, they were able to assess whether couples of genes were mutated in the same patient at the same time (co-occurrent) or significantly rarely found

together in the same patient (mutually exclusive). They uncovered the presence of co-occurrence patterns between NPM1 and FLT3 and NPM1 and DNMT3A; on the other hand, NPM1 and FLT3 resulted mutually exclusive with RUNX1 and TP53. All these results can support the definition of a putative mechanism for AML development.

Figure 1.6: Functional categories for mutations identified in AML patients. For every patient the box is filled in correspondence of a mutation. Genes are grouped in the 9 classes detected by HotNet, cytogenetic risk and types of mutation are explicated by the colour of the box.

(Adapted from TCGA, 2012[44])



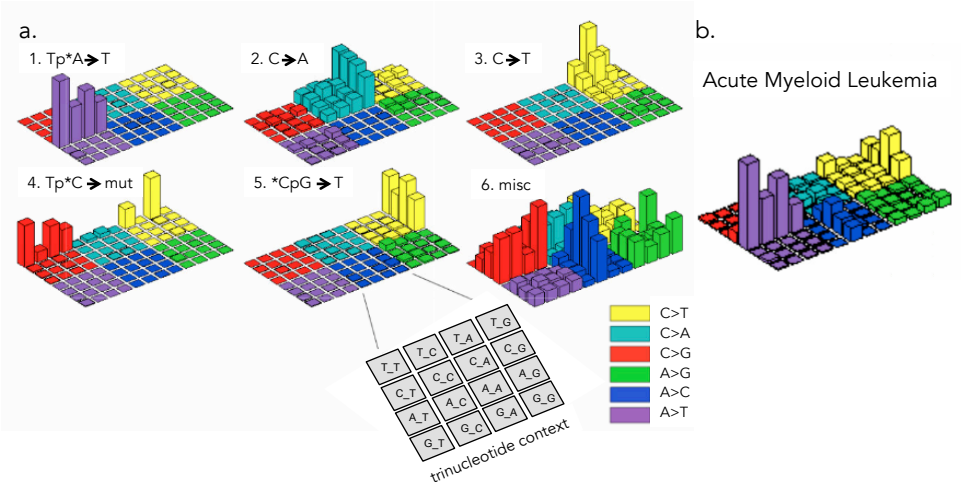
1.3.2.3 Identification of mutational spectra

Understanding the origin of the mutations is valuable for many reasons: it can pinpoint the patients that have higher risk of developing the disease and

anticipate diagnosis, advance new potential therapeutic targets and enforce the promotion of healthier life styles. Several environmental factors can foster the appearance of new mutations; the use of reverse engineering approaches can guide to deduct the causal relationship between mutations and environmental factors. Analysing the mutations in a trinucleotide context is a step forward analysis that comes after mutation identification. It considers the mutations and the flanking basis as a single unit and recognizes typical patterns of mutations related to a specific tumour; consequently it is able to suggest a possible mechanism for mutation appearance. The mutational spectrum for a pathology is represented as a three-dimensional bar plot (Figure 1.7) in which the bars are disposed as six rectangles coloured on the basis of the point mutation considered; each rectangle contains sixteen bars correspondent to the possible trinucleotide contexts for that point mutation. The height of the bar represents the number of mutations of that type. The NMF algorithm divides all the TCGA patients in six factors characterized by peculiar spectra of the mutational landscape described. Tp*A -> T transversion dominates the AML mutational spectrum; this type of mutation is mainly found in leukaemias (AMLs and CLLs), but its interpretation is not yet clear.

Figure 1.7: Mutational spectra. a. Description of the composition of mutational spectra: every rectangle on the plane represents a mutation in a trinucleotide context. The six different colours are associated to the six possible point mutations and the group of boxes of the same colour are linked to all the possible trinucleotide contexts for that point mutation. Analysing all the tumours in the TCGA cohort, Lawrence et al. identified six typical mutational spectra behaviours; b. AML mutational spectrum is dominated by the Tp*A -> T base change

(Adapted from Lawrence et al., Nature 2013)



1.4 AML risk increases with age

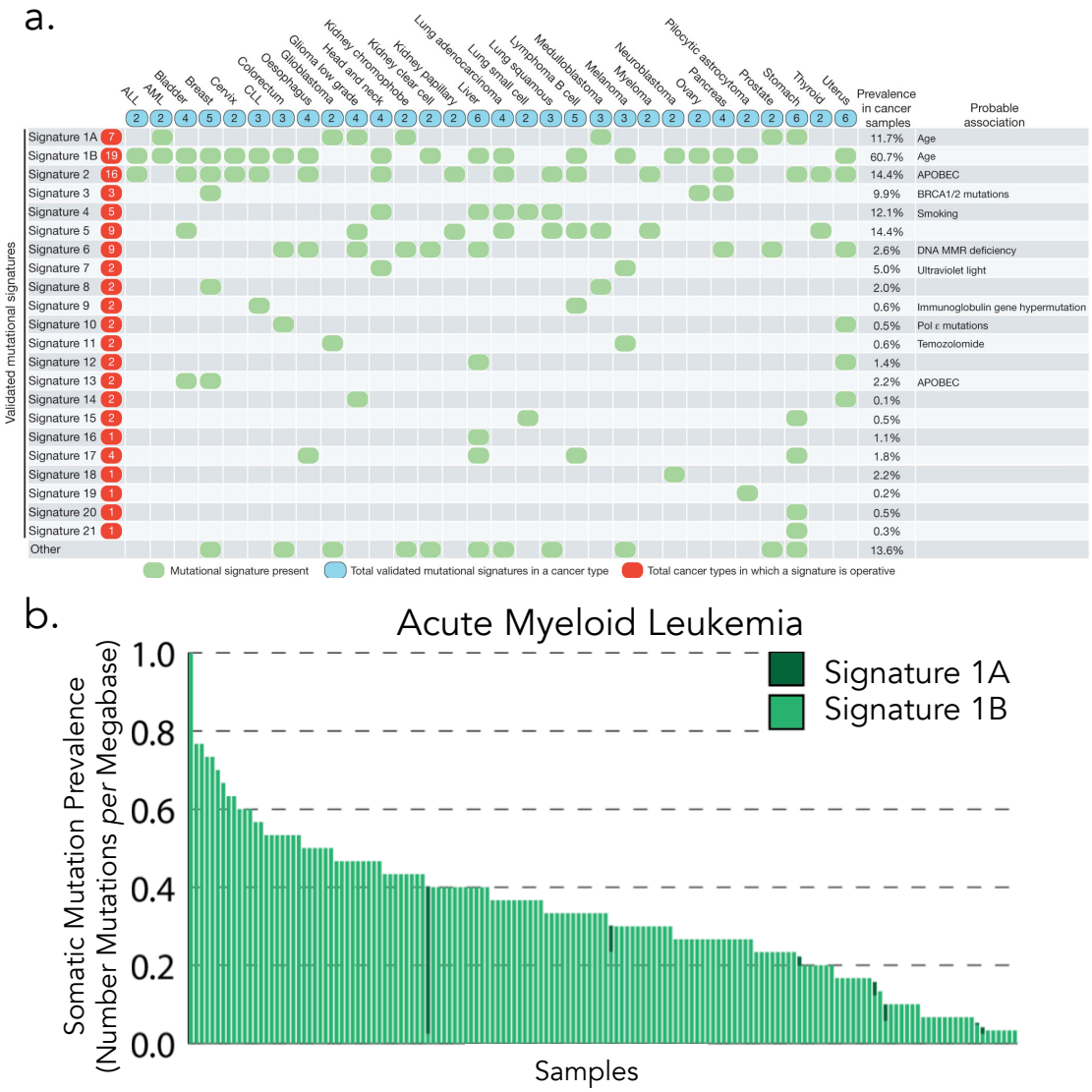
The mutational spectrum analysis described in paragraph 1.3.2.3, grouping mutations together by base change and the genomic context in which the mutation occurs, gives a first input on the possible mutational process that led to the development of the disease. The consequent intriguing question would be whether there exist combinations of mutations that can be associated to a specific mutational process (mutational signatures). Alexandrov et al.[49] developed an algorithm able to extract the mutational signatures from groups of patients affected by the same cancer. This algorithm considered mutation types in a trinucleotide context and found the solution in which the smallest number of signatures better explains the observed portion of mutations. Subsequently, the

authors used hierarchical clustering to put together the same signature coming from different cancer types and end up with 27 signatures for the 30 cancer types analysed. They were also able to associate some of these mutational signatures to possible mutational mechanisms[49]. As an example, signatures 1A and 1B (Figure 1.8.a) are present in the majority of cancers and have been both related to aging, because their abundance correlate significantly with the age at diagnosis in many cancer types. The strong correlation with age suggests that the mutations belonging to these signatures are naturally acquired during the lifetime of an individual. Mutation rate heterogeneity among individuals suffering of the same cancer type can be due to diverse environmental factor risks (e.g. exposure to carcinogens) or to the acquirement of mutations affecting repair mechanism or other pathways that promotes mutation accumulation before/after cancer initiation (Figure 1.8.b). Because both signatures 1A and 1B are associated with age, they are mutually exclusive between cancer types. However, the former is always found in copious groups of patients and the latter in smaller groups, suggesting that they are just two version of the same signature.

This type of analysis applied to AML showed that AML patients present only signature 1 in both forms (A and B). The majority of the patients show signature 1B probably due to the fact that AML patients in this study were 200 (Figure 1.8.b).

Figure 1.8: Mutational signatures identified across human cancer types. (a) for every signature identified is reported in which (and in how many) cancer types it was found as significantly enriched and the probable association with mutation causes and mechanisms; (b) for every AML patient is reported the prevalence of mutational signatures as the number of mutations for that signature *per* Megabase.

(Adapted from LB Alexandrov *et al.*[49])

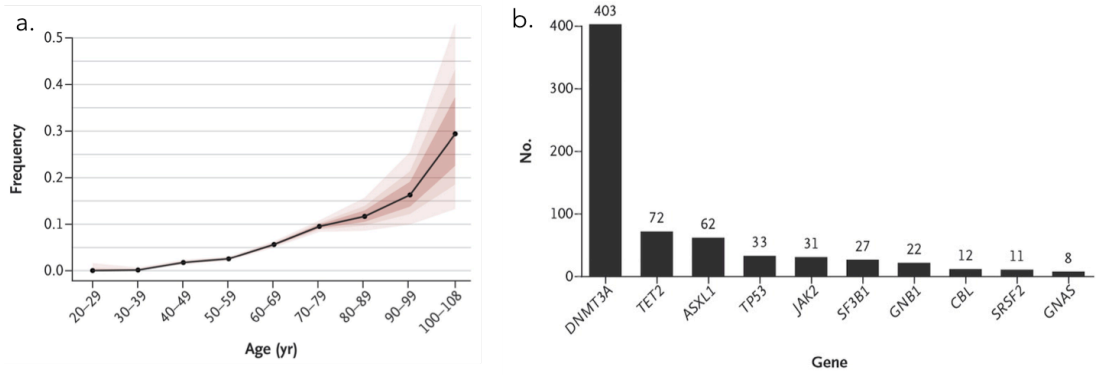


These results are in line with the fact that AML is typically a cancer of advanced age, however unexpectedly also healthy elderly individuals can present signs of clonal haematopoiesis. Early studies detected clonal haematopoiesis events in 23.1% of normal elderly women (mean age 76) through the observation of X chromosome inactivation patterns using HUMARA assay [50]. The authors hypothesized three possible scenarios for these results: the slow selection of

differences related to the X chromosome, a stem cell depletion that leads to clonal dominance over time or a clonal advantage given by the mutations. The last theory was confirmed by the detection of recurrent TET2 somatic mutations in normal elderly individuals[51]. Also copy number alterations resulted to have an increased frequency after 75 years: from 0.23% before 50 years to 1.91% between 75 and 79 years[52]. Moreover, it has been shown that the blood of normal individuals with no signs of haematological disorders present one or two somatic mutations in genes already described as drivers in haematological malignancies. In particular, the most frequently mutated genes are DNMT3A (which is largely the most represented and has been found mutated also in the blood of remission patients up to 8 years after initial AML diagnosis)[53], TET2 and ASXL1[54,55] (Figure 1.9.b). Figure 1.9.a reports the frequency of somatic mutated individuals divided in groups according to their age. Since the 17'182 normal samples analysed came from a cohort of WES samples collected for other scope than haematological malignancies, it turns out clearly the impact of aging, and therefore of time, on the occurrence of somatic mutations. Somatic mutations in healthy peripheral blood increases tenfold comparing people under 50 and over 65 years old and, in parallel, also the risk for haematological cancer increases (hazard ratio, 11.1; 95% confidence interval [CI], 3.9 to 32.6 and hazard ratio, 12.9; 95% confidence interval, 5.8 to 28.7 in two different studies[54,55]).

Figure 1.9: Elderly normal individuals present and augmented rate of mutations in the peripheral blood cells. a. Frequency of somatic mutations is reported for every group of age from a cohort of 17,182 normal samples, the red area refers to the 50th, 75th and 90th percentiles; b. Number of mutations identified in the ten most mutated genes in normal individuals.

(Adapted from Jaiswal et al.[56])



1.5 Leukaemogenesis

There are many questions regarding leukaemogenesis that at present are still unanswered as the characterization of the cell of origin of AML, whether there is a predefined order for the appearance of the mutations that lead to the development of the leukaemia, the possibility that pre-leukemic cells persist after treatment and guide the successive relapse. All these questions have an impressive impact on the clinic, the search for new treatments and the outcome for patients.

There is little information about the preleukemic phases of AML, because, of course, typically AML is diagnosed after full evolution of the disease. For this reason it is difficult to trace the exact consecution of changes that lead a normal cell to become leukaemogenic. However, there are some AML patient subtypes for which some developmental information has been collected. Myelodisplastic Syndrome (MDS) is a malignancy that often precedes AML: at the cellular level

the bone marrow becomes clonal but the cells maintain their ability to differentiate[57] and have a high rate of apoptosis[58]. In some patients affected by MDS, these two latter cellular characteristics are lost and the disease evolves in acute leukaemia. The mechanisms at the basis of the development of MDS have not been identified yet and there is not a common mutation to all cases that can be pointed out as the responsible for this pathology. Furthermore, patients present a phenotypic heterogeneity that may have a genetic origin, although there is no clear genetic profile associated with the different phenotypic categories. Indeed, MDS patients show mutations in many AML genes that are associated to poor prognosis: those belonging to the chromatin regulation compartment as DNMT3A, TET2, IDH1 and IDH2 involved in the CpG island methylation of promoter regions; ASXL1 and EZH2 that are members of the polycomb family and act modifying the histone proteins; oncogenes and tumour suppressors like RUNX1, ETV6, TP53, EVI1, JAK2. They often present uniparental disomy (10-15% of patients) that consists in the acquisition of both copies of a chromosome (or part of a chromosome) from the same parent, and in many cases small insertions and deletions are present. Finally the genetic lesions that mostly characterize MDS patients are spliceosome mutations (e.g. SF3B1, SRSF2, U2AF1, ZRSR2) that are present in 85% of the cases. In general, these mutations are mutually exclusive and may play a role in the splicing of crucial genes like TET2, RUNX1 or act cooperatively with other genes as DNMT3A that often co-occurres with SF3B1[40,59].

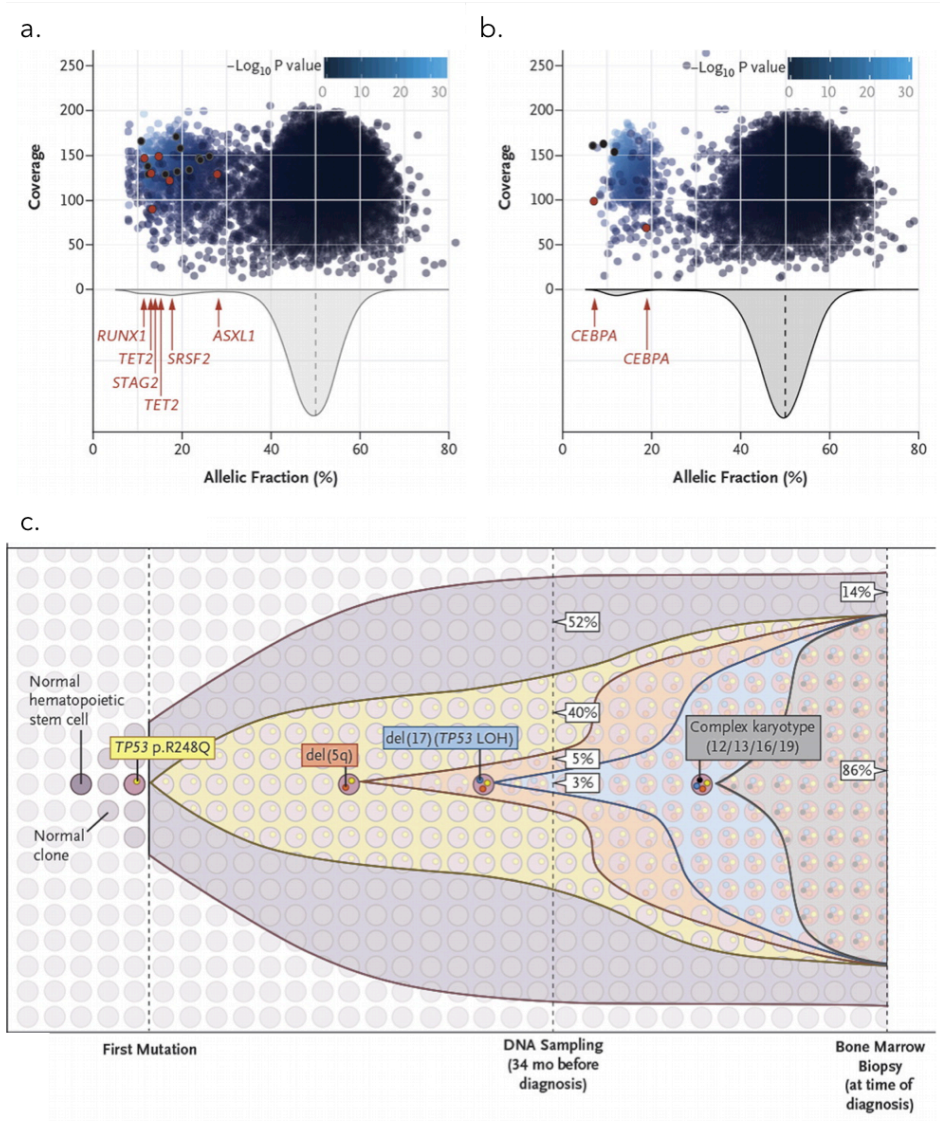
Observation of the MDS patients provides a landscape for the emergence of the

leukaemia but it is possible that these mechanisms do not retrace the AML development in not myelodisplastic patients. Elderly normal individuals with clonal haematopoiesis that eventually evolved in AML are the source of additional information on the genesis of the pathology[55]. Up to date, three cases have been reported and two of them developed leukaemia very rapidly: two months after clonal haematopoiesis was observed, they were diagnosed respectively of MDS and AML. The authors reanalysed by WGS the first blood sample collected and identified many AML mutations in the first patient (Figure 1.10.a) at frequencies consistent with the presence of a clone. The mutation with highest frequency was ASXL1, considered the putative founder of this clone; other known drivers identified were RUNX1, TET2, STAG2, SRSF2, all mutations also identified in MDS patients. The second patient (Figure 1.10.b) had two different mutations on CEBPA that were found also in the bone marrow at diagnosis (together with three putatively passenger somatic mutations) at a slightly augmented frequency. The prognosis for this patient was favourable and he achieved complete remission after treatment. For the third patient, the authors were able to hypothesize the sequence of genetic lesions that predisposed the emergence, 34 months after the first sampling, of AML (Figure 1.10.c). In this case they identified a TP53 mutation, present at 86% of allele fraction at diagnosis. This mutation was present at 23% in the precedent blood sample tested. Thanks to the concurrent presence of a number of chromosomal rearrangements, they were able to put in a temporal sequence the changes described in the Figure below (1.10): the tumour arose from a subclone that was present at 3% frequency in the first sample examined

and that gained additional aberrations, becoming more aggressive.

Figure 1.10: The evolution of clonal haematopoiesis in 3 individuals that developed haematological malignancies after being sequenced as normal elderly. (a,b) For patient 1 and patient 2 are shown the VAFs of the mutations identified in function of the sequencing coverage. The –log10 p-value is associated to the probability of each mutation of having a VAF lower than 50% (binomial test). Red dots indicate the variants that hit known driver genes in leukaemia. Panel c. shows the reconstructed evolution of the rearrangements occurrence in patient 3 as indicated by WGS. Percentages have been estimated based VAFs, the colours are associated to the genetic lesions depicted in the box at the beginning of the shading and its expansion.

(Adapted from Genovese et al.[55])

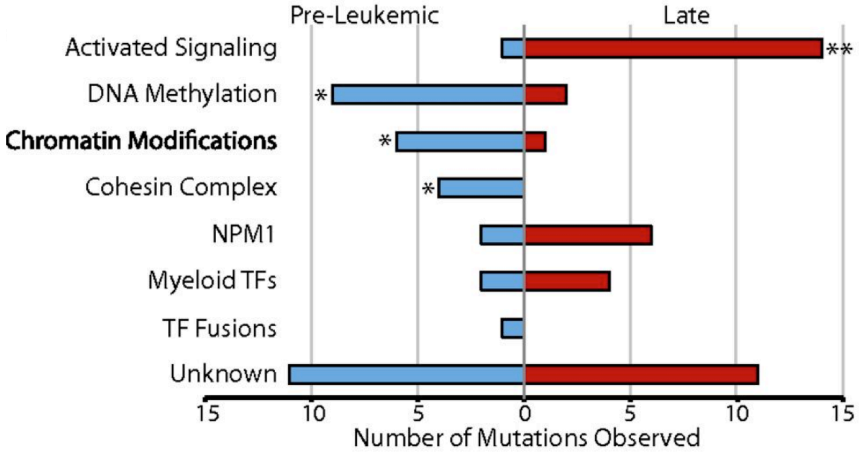


Majety's group hypothesized that the leukaemia cell of origin engenders from a HSC. Self-renewal capacity and longer lifespan would allow accumulating mutations circumventing the low rate of spontaneous mutations. In fact, they

were able to identify through WES, leukaemia associated mutations in a population of HSCs with normal activity isolated from AML patients. Mutations in common between the HSCs and the frank leukaemia hit genes like NPM1, TET2, SMC1A and CTCF. Furthermore, in five cases, the authors were able to find some mutations of the frank leukaemia but not all, indicating that the cell of origin of the leukaemia started from the HSC compartment and accumulated additional mutations[60]. In a successive study[61], the authors separated HSCs on the basis of the expressed surface markers and divided the leukaemia mutations in: i) pre-leukemic mutations, mutations present in the HSC population at VAF higher than 1%; ii) late mutations, mutations absent in the HSC population. They, then, stratified the patients by functional categories, based on the function of the mutated genes, and uncovered that some categories were preferentially mutated in the pre-leukemic phase and some in the late phase (Figure 1.11). In agreement with the results already obtained from mutational analysis of MDS and normal elderly individuals, the authors found, in the first phase of the disease, an implication for "landscaping"[61] genes, that include DNA methylation, chromatin modification and cohesion complex genes; and for late mutations, mutations that affect signalling genes that promote the progression to overt leukaemia.

Figure 1.11: Mutations of 16 patients and their occurrence in early or late phase of AML development stratified by categories. Mutations that were already detectable in the HSCs of the AML patient are considered pre-leukemic, those that do not fulfil these requirements are considered late. Mutations were divided in subgroups as described in TCGA paper[44].

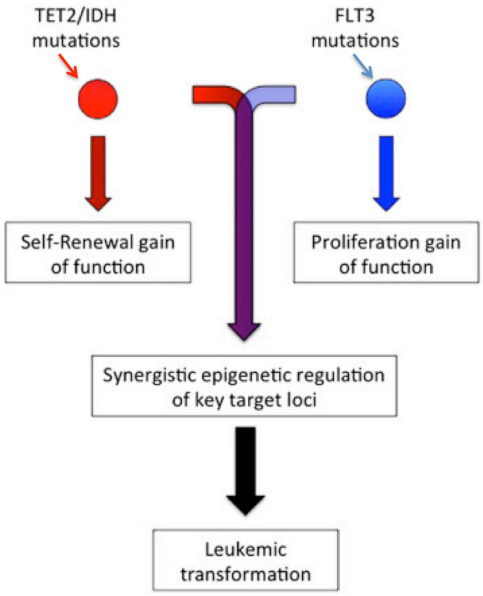
(Adapted from M. Ryan Corces-Zimmerman et al.[61])



In particular, landscaping mutations are thought to prepare the HSC to clonal expansion conferring a competitive advantage as it has been demonstrated for TET2[38], DNMT3A[62] and IDH1[37]. Therefore, the leukemic transformation appears to be the combined effect of two successive alterations that confer first a self-renewal gain of function followed by a proliferation gain of function (Figure 1.12).

Figure 1.12: Schematic representation of the synergistic effect of landscaping mutations and activated signalling mutations necessary for leukemic transformation. The cooperation of these two types of mutations enhances the capacity of a HSC to become leukaemogenic.

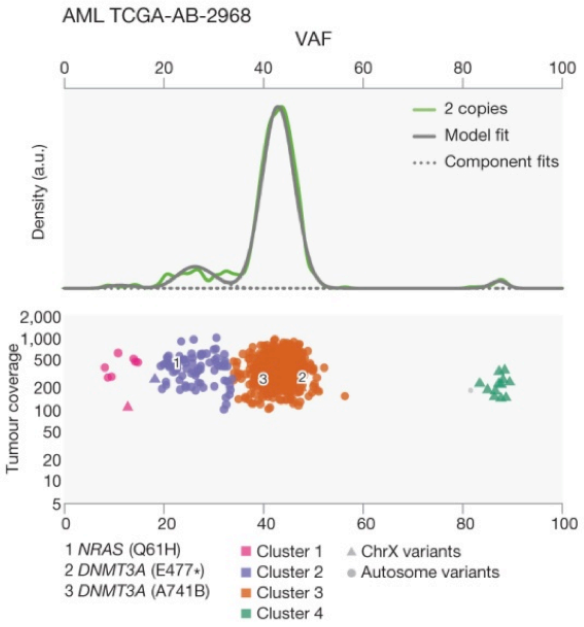
(Adapted from Sykes et al.[63])



1.6 The complexity of clonal architecture

In paragraph 1.5 we described the emerging evidence on the genesis of AML, describing it as a linear process. However, that is an oversimplification that allowed us to explain the concepts on the process underlying. Now we can examine in depth the context that subtend the events previously described. Indeed, in the last decade, with the advent of NGS, the complexity of clonal architecture of cancer is becoming more and more evident. In fact, the composition of the cancer population, instead of being formed from a single clone accumulating successive mutations, is more likely composed of multiple clones, characterized by different genetic compositions that exist together at different frequencies in the tumour population. An example of this behaviour is reported in Figure 1.13 in which it is possible to identify 4 different clones in the AML population. The authors filtered out CNVs, keeping only regions with two chromosomal copies and, afterwards, clustered the mutations with similar VAF to obtain the possible mutational composition of distinct clones. DNMT3A mutations, considered the initiating mutations for this patient, appear in the clone at highest frequency (considering that frequencies of X-linked mutations in males have doubled VAF). This AML clone can be considered the dominant clone, but there are at least two other subclones present in the sample at lower frequencies.

Figure 1.13: The clonal composition of a leukemic population in a patient. The plot is divided in an upper panel, that displays the density of mutations at a specific VAF in the sample, and a lower panel in which the VAF is connected to the coverage at each position. The colours represent distinct subclones (clusters), as predicted by the algorithm SciClone. Numbers highlight the position of driver mutations in the clones.
(Adapted from C. Kandoth *et al.*[64])



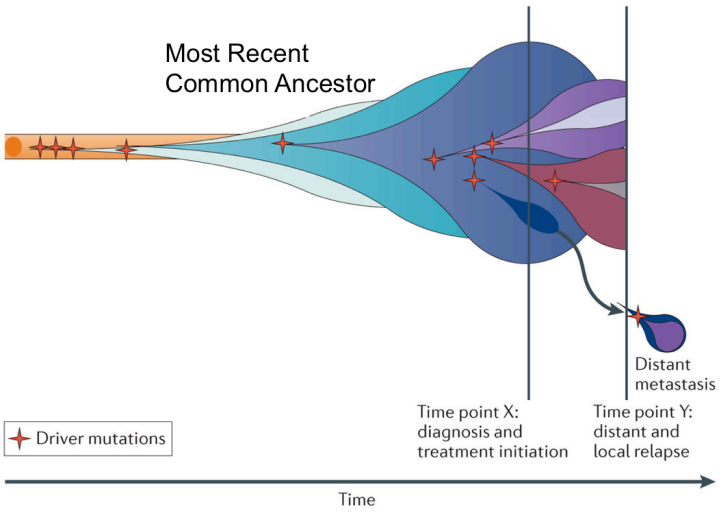
Indeed, heterogeneity has been observed across many cancer types and there seems to be a general trend for clonality in cancer: tumours often show the presence of more than a single clone *per* patient and in a study of 1'165 patients analysed across 12 cancer types, 86% had at least two clones. Furthermore, the number of sub-clones forming the tumour seems to correlate to the mortality risk for the patients. When the number of subclones is greater than 2, the hazard ratio is 1.49 (95% CI: 1.20–1.87)[65], comparing them to tumours formed only by one or two clones. Interestingly, there seems to be an upper limit for the number of clones in a tumour: when the clones are more than four, genomic instability likely becomes problematic for the tumour itself, reducing the risk for the patient.

The forces that interplay giving rise to intra-tumour heterogeneity are somatic mutations, natural selection and adaptation to the tumoural microenvironment.

Environments (like the primary tumour site, the organs target of metastasis or the exposure to carcinogens, as UV light, tobacco smoke or chemotherapy) drive the selection of some clones in favour of others. In Figure 1.14 is schematized the effect of the environment on cancer evolution at different stages. Every new mutation in a cell enters in a picture that was already designed and its emergence can be neutral, give an advantage or be deleterious for that cell [66]. The effect of addition of a mutation to a cell is called epistasis and the impact of epistasis on tumour evolution is confirmed by the presence of co-occurrence and mutual exclusivity of mutations in cancer genes (as already described in paragraph 1.3.2.2)[44]. Furthermore, cancer evolution does not always occur stepwise but catastrophic events may take place, changing dramatically the genomic landscape of a cell (e.g. chromotripsis).

Figure 1.14: The evolution of cancer genome. This scheme reports a complex hierarchical composition of cancer that arises from the emergence of many driver mutations in different tumour subclones and can generate relapses and metastasis genetically different from the dominant clone in the primary tumour.

(Adapted from Yates et al.[66])



Many approaches, both biological and mathematical, have been used in the last years to uncover the genetics of heterogeneity in cancer and the main categories of these strategies are discussed below.

1.6.1 Multi sampling

Multi sampling consists in the analysis and comparison of repetitive sampling from the same patient. Samples can be temporally distinct, as, for example, the diagnosis and relapse samples of a patient, or spatially distinct, like the parental tumour and the metastases of the same neoplasm or sampling of different portions of the same tumour mass. Recently, many groups have started to combine these two approaches collecting samples both temporally and spatially distinct from each other and analysing them together.

An example of multiple sampling is reported in the study of Schramm et al.[67], in which the mutational analysis of a primary neuroblastoma was compared to the analysis of its five relapse and metastasis samples. The authors performed WES analysis of all the samples and, subsequently, after filtering the positions that reached the minimum quality requirements in every sample, they computed a table containing all the alleles that showed a different base (e.g. nucleotide variant) from the others in at least one sample. Using this table, they simply calculated the Hamming distance (i.e. number of substitutions needed to obtain one sequence from the other), as the number of positions in which they differed,

between each couple of samples. After using neighbour-joining algorithm[68] to construct the tree of relationships among the samples, they uncovered that every relapse contained private mutations and that there were at least two different clones at the origin of the very heterogeneous relapse and metastasis samples.

There is a flourishing publication of methods aiming at the reconstruction of clonal composition starting from the variations identified in multiple samples from the same patient[69–72]. Despite the mathematical models at the basis of these methods are sometimes very different, all of them use VAFs, corrected for the copy number prevalence at the site of the mutation, to group variants into clusters in the sample. We describe in details these different methods in the Materials and Methods section (paragraph 3.6).

1.6.2 Single-cell sequencing

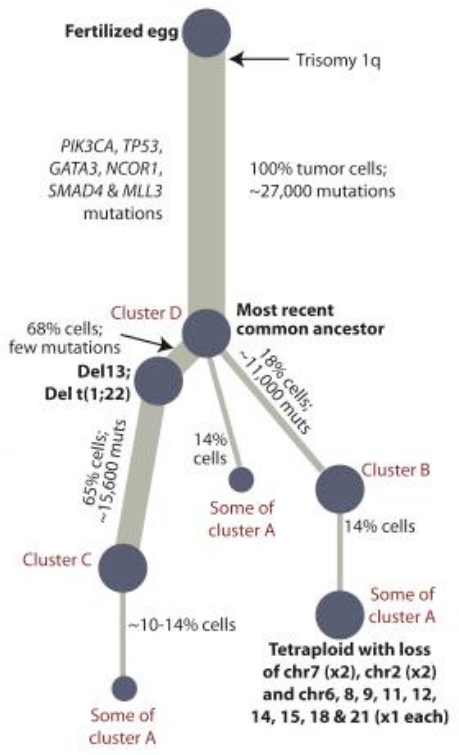
Single cell sequencing is the more precise method to infer clonal composition of tumours because it effectively permits to determine which variants, indeed, belong to the same cells. However, this method is costly, has some bias that still need optimization and, for the time being, is not ready for high throughput analyses. An example of a study using single cell sequencing in two breast cancers to reconstruct the subtended tumour populations was performed by Navin et al. in 2011[73]. The authors analysed 100 single cells from two tumours, one monogenomic and one polygenomic; in both cases they uncovered the presence of a consistent subpopulation of genetically diverse cells that, for the

case in which they disposed also of a metastatic sample, were not identified at the metastatic sites.

1.6.3 Mathematical and statistical models

In order to deal with the issue of the resolution of an admixture of cells in a single sample, also computational approaches can be exploited that allow to determine the genetic makeup of the cell subpopulation and their prevalence in the sample. Greenman et al.[74] developed a method that employs point mutations and genomic rearrangements to reconstruct the sequence of events that took place in a sample, building a sort of historical reconstruction with the use of the graph theory. In 2012 Nik-Zainal et al.[75] used this algorithm to reconstruct "the life history" of 21 breast cancers. In their cohort of patients, they were always able to identify a dominant clone representing about 50% of the tumour cells. The remaining population was formed by a great number of low-frequency subclones harbouring hundreds to thousands of mutations (see Figure 1.15 for an example). This finding enabled the authors to propose the existence of a quiescent reservoir of cells that are potentially capable of repopulating the tumour after the acquisition of new proliferative advantaging mutations.

Figure 1.15: The phylogenetic tree reconstruction for a breast cancer patient. The authors reported the alterations characterizing each clone and sub clone and the abundance of the cells in the tumour population (thickness of the grey lines).
(Adapted from Nik-Zainal et al.[76])



1.7 State of the art of treatment in AML and determination of remission in patients

With the exception of APL, the specific M3 subtype of AML, which has a distinct mechanism of leukaemogenesis and for which has been developed a successful molecular treatment, treatment of the other AML subtypes is general and is consolidated, with no great changes in the administered drugs since many years. The treatment is divided into two main phases: the first aims at the rapid eradication of the AML blasts and induction of the remission of the leukemic state (remission induction or, simply, induction therapy); the second aims to prevent the relapse of the disease (consolidation or post-remission therapy).

Since leukaemia is a very fast growing tumour, the induction therapy needs to be performed immediately after diagnosis and physicians, generally, choose the strongest treatment the patient is able to sustain, in order to have a good probability to eradicate the tumour. The general, treatment for the induction therapy is known as "3+7" and involves three days of anthracycline administration (daunorubicin, idarubicin or anthracenedione mitoxantrone) and seven days of cytarabine. Complete remission (CR) is achieved in 60-80% of young patients and in 40-60% of elderly patients[77]. Many other drugs and combination of drugs have been used through the years to test their possible advance in treatment outcome, however, based on the evaluation of the risks given by toxicity, the CR rate and the overall survival, the "3+7" is still the best compromise. After the induction therapy the leukaemia population is reduced under the cytogenetically detectable threshold (~10^9 cells); however some leukemic cells can persist after treatment and, without a supplementary treatment aimed at maintaining the remission state, these cells can eventually lead to relapse. This second phase of treatment is called consolidation therapy and comprehends additional drugs and/or bone marrow transplantation (or hematopoietic stem cell transplantation, HSCT). This second phase of the treatment is more compliant with patient's and AML's specific characteristics. It has been demonstrated that four cycles of high doses of cytarabine (3 g/m^2 per q12h on days 1, 3, 5) give better results than lower doses[78]; similar response should be achieved with other chemotherapeutic agents at high doses. Autologous HSCT has effects similar to chemotherapy and is recommended only for high risk cytogenetic patients, allogeneic HSCT is

associated to low relapse rate both in intermediate and high cytogenetic risk patients[79]. The positive effect of these treatments is given by the graft-versus-leukaemia effect[80] that is an anti-tumour response arising only after transplantation.

Different treatment is reserved for older patients (≥ 60 years) for which the standard therapy is more often toxic and have an increased relapse risk; with standard "3+7" treatment their life expectation is 8 to 12 months[81]. It is possible that the characteristics of age-related leukaemia, already elucidated in paragraph 1.4, impart a stronger resistance to the disease and results into poorer outcomes. Activation of the RAS, SRC and TNF pathways may play a role in these events.[82] Therefore, suggested treatment for patients between 60 and 74 years of age has a "3+7" walk but envisages cytarabine at reduced doses compared to standard treatment, doses that need to be adapted to the patient characteristics. The same rules applies to the consolidation therapy: only rare cases benefit from dose escalation after the first chemotherapy and, usually, allogeneic HSCT offers the best results.[79] After 75 years of age, the choice of the treatment has to be taken together with the patient because low doses of cytarabine can be toxic, resulting in a 30 days mortality rate of 26%. Furthermore, cytogenetics and AML type have a greater impact on the response to treatment in this age category.[79]

The possible outcomes after standard treatment are diverse, because patients can be respondent, partially respondent or not respondent to therapy, according to achievement of complete remission (CR) after induction therapy. Achievement of complete remission is defined by the following criteria:

- percentage of blasts in the bone marrow below 5;
- total absence of blasts with Auer rods;
- values for neutrophil and platelet count in the normal range: > 1 x 10^9/l and 100 x 10^9/l, respectively;
- independence from red cell transfusions;
- absence of extramedullary disease.[83]

The remission state of a patient can be established at different levels, based on the type of technology used to determine it:

- Morphologic: comprehends all the parameters discussed above except for counts of neutrophils and platelets;
- Cytogenetic: applicable only in AML cases presenting cytogenetic abnormalities, consists in the return to a normal karyotype;
- Molecular: different molecular markers can be used for testing and, in general, remission is achieved when the molecular marker tested at diagnosis is below the level of detectability.

In some cases CR is achieved but the recovery is not complete (incomplete Complete Remission, CRi), because the patients present low neutrophil or platelet counts. Alternatively, the treatment may give partial results, with a decrease of the blast percentage insufficient to achieve CR (partial response), or no results at all. In this last case, the disease is defined as resistant because the leukemic cells are not affected by the chemotherapy.

Despite the high rates of CR achieved after treatment, the number of patient that will result cured after induction and consolidation therapy is very low (~12%).[84] In

the next paragraph, we discuss the characteristics of relapsing leukaemias and the possible mechanisms that underlay treatment failure.

1.8 Relapsing AML

Relapsing AML can arise from months to years after the first CR, and the first three years after CR are particularly crucial, because, beyond this period of time, the risk of relapse reduces steeply.[84] The relapse free percentage at 6 years for patients, that already reached three years of disease free survival, was estimated to be ~86% on an American cohort.[85] However, the percentage of relapsing patients is very high and understanding which patients have a higher probability to relapse would have a great impact on the clinics.

Several markers have been identified that help the prediction of outcome at diagnosis. They are mainly based on cytogenetics, clinical information, clinical history or molecular factors. Age, as already discussed, is an important prognostic factor; old patients are more difficult to treat and relapse is more frequent in old than in younger individuals.[86]

Cytogenetic plays an important role in the prediction of the treatment outcome and the risk stratification, already described in paragraph 1.3. Grimwade et al.[16] examined the rate of relapse at five years in a group of 1'612 patients; the overall percentage of relapse is 49%, but its range changes substantially in the three risk groups, as elucidated in Table 1.4.

Table 1.4: The influence of cytogenetically defined risk categories on relapse risk at 5 years.

Risk category	Cytogenetic abnormalities	Relapse risk at 5 years (range)
Favourable	t(15;17), t(8;21), inv(16)	29% - 42%
Intermediate	NK, +8, 11q23, +21, del(7q), del(9q), +22, other numerical or structural abnormalities	39% - 60%
Adverse	Complex karyotype, -7, abn(3q), del(5q), -5	68% - 90%

Independently from the cytogenetic risk groups, the presence of the FLT3 ITD mutation is an important predictor for relapse in AML patients. Its presence is, in fact, strongly associated with increased relapse risk, adverse disease free survival and overall survival.[87]

In contrast, NPM1 variants are generally associated to the favourable risk class, because the patients harbouring this mutation generally respond well to chemotherapy. However, the co-occurrence of NPM1 and FLT3 mutations weakens the respondent phenotype and shifts the response to therapy towards the adverse prognosis observed in patients harbouring the FLT3 ITD mutation.[88] Considering their impact on disease outcome and response to therapy, both FLT3 and NPM1 are used as fundamental markers for CR and their presence is tested during the follow up of the patients in order to detect the possible presence of residual tumour cells, defined as minimal residual disease.

All of these markers show strong association with relapse free survival but they, even considered all together, are neither necessary nor sufficient for a confident forecast. For this reason, it would be of great clinical impact to identify new

markers that may allow the prediction of treatment outcome in AML patients at time of diagnosis.

The treatment options for relapsing leukaemia mainly depend on the time in which relapse occurs: if relapse occurs within one year after the first CR, it is highly probable that the AML is resistant to therapy and the general suggestions is to consider the use of experimental drugs, followed by a HSCT, in case remission is accomplished; if the relapse occurs later, the first treatment of choice is the combination of drugs like daunorubicin, idarubicin and cytarabine[89].

The mechanism that causes relapse is known only for specific cases but for the majority of the patients it has not been elucidated yet; it might be possible that genetic abnormalities escaping conventional remission assessment techniques (i.e. minimal residual disease) can result in the recurrence of the disease months later.

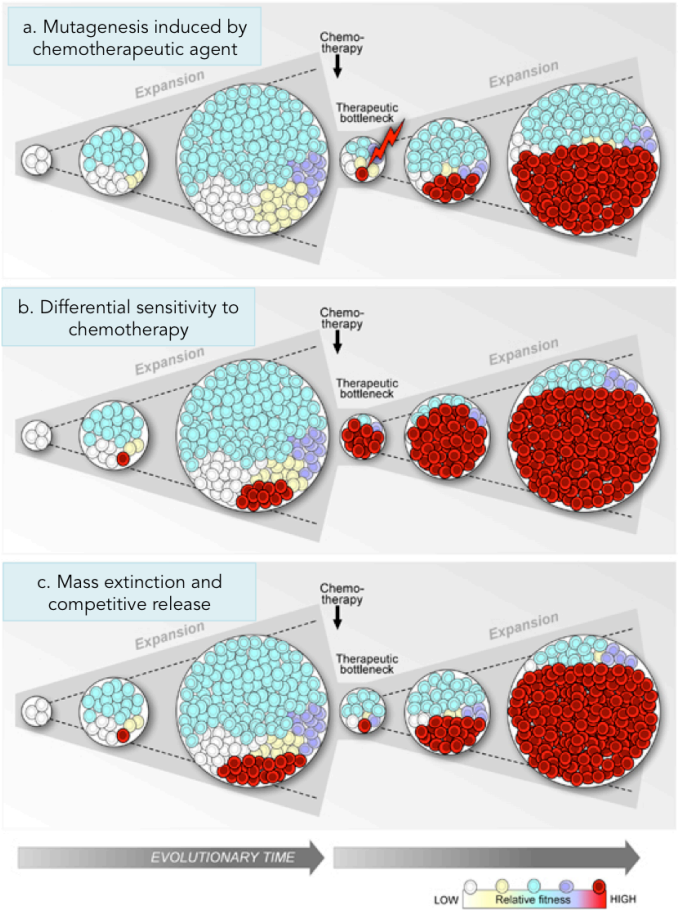
1.9 Clonal evolution in AML

After this excursus on the clinical and biological presentation of relapsing AML, it sounds clear that it is of paramount importance to elucidate the mechanisms and the causes of relapse of the disease. Therefore, probably, the path that will guide us in understanding the molecular players that lead to relapse requires looking at relapse from an evolutionary point of view. Also the role of chemotherapy in induction of new mutations and in clonal selection needs to be further elucidated. The mutagenic effect of cytotoxic agents used to kill the tumour cells, in some

cases, can induce new mutations that provide selective advantage to the cancer, posing the basis for relapse (Figure 1.16.a). Indeed, chemotherapy may act as a selective force and, after treatment, clones resistant to the drug can expand (Figure 1.16.b). Alternatively, if the drug has the same effect on all main subclones, the empty niche may become the breeding ground for pre-existing clones or subclones that best fit to the new environment (Figure 1.16.c).[90] These possible scenarios highlight the fundamental role played by tumour heterogeneity in AMLs: the presence of many subclones in the same patient and the possible presence of a reservoir of highly mutated cells enhance the chances for tumour regrowth and chemotherapy escape.

Figure 1.16: possible AML evolutionary scenarios in case of unsuccessful chemotherapy. a. Many drugs used for chemotherapy act through the induction of new mutations in the cells, a drawback of this mechanism is the possibility to induce mutation favourable for the progress of the disease; b. The sensitivity to chemotherapy can be modulated by many genetic and phenotypical features; for this reason cancer subpopulations can react differently to treatment; this can result in the emergence at relapse of populations that were subclonal in the primary tumour; c. When response to therapy is similar for all the clonal subpopulations the new environment is exposed for the expansion of the fittest clone, irrespectively of the mutations appearance time.

(Adapted from Landau et al. 2014[90])

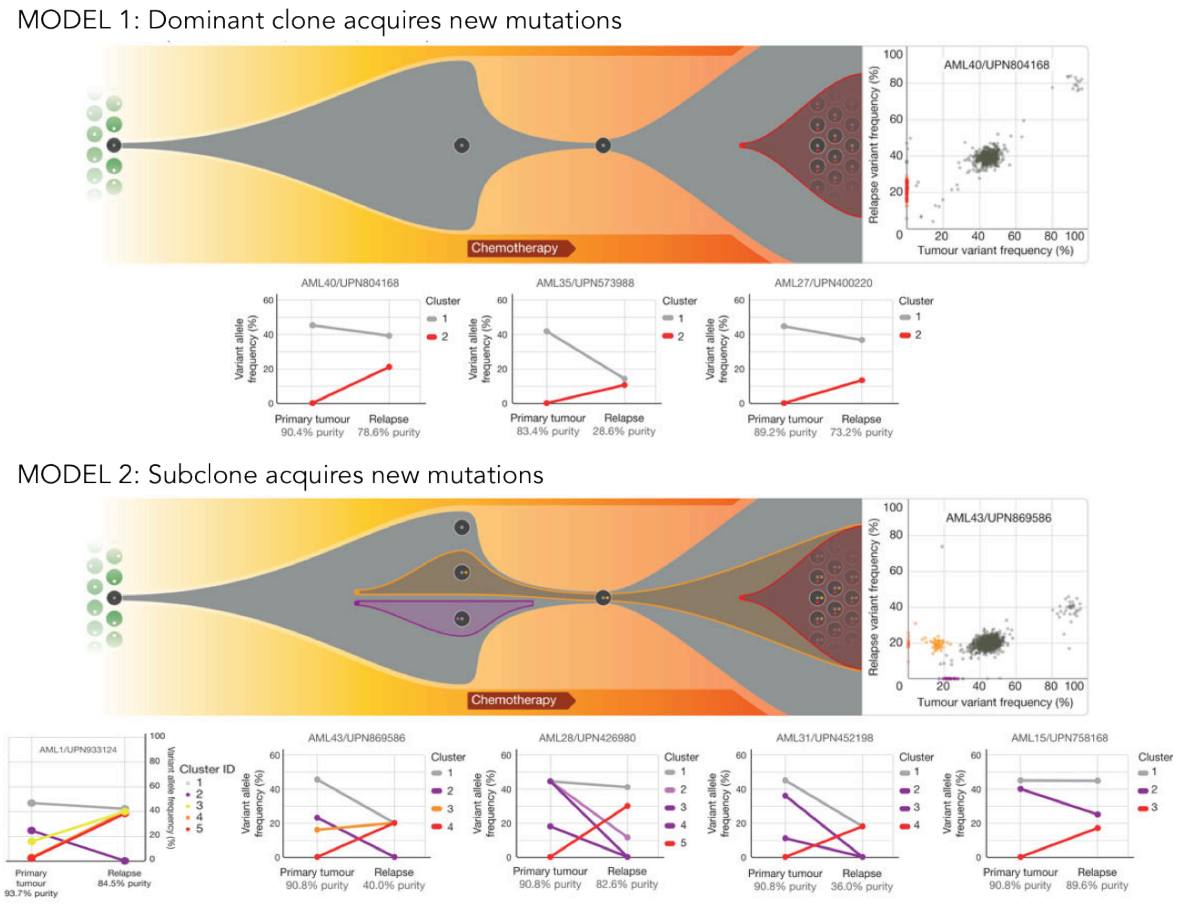


In this context, many studies compared patient's DNA at exordium and relapse in order to identify possible markers and mechanisms for relapse emergence. In 2012, Ding et al.[91] analysed the whole genome sequence (WGS) of 8 patients and described two alternative behaviours for their relapse: in 3 cases the dominant clone gained additional mutations at relapse, in 5 cases the relapse arose from a subclone already present at exordium that successively accumulated additional mutations (Figure 1.17). For model 1 the authors suggest that the patients could

be inadequately treated or the founder clone already contained a mutation conferring resistance, for model 2 they suggest the presence of resistant mutations or the induction of new crucial mutations by cytotoxic agents in a specific subclone.

Figure 1.17: Two scenarios of evolution of the disease that lead to relapse in AML patients. In the first model the relapse arises from the dominant clone, which is not killed by chemotherapy and further evolved in the relapse. Clusters were identified grouping mutations based on their VAF at exordium and relapse, as shown in the graph at the right of the model. In the three patients presenting this type of relapse, we observe the presence of just two clones: the grey clone was already present at exordium and survived chemotherapy; the second clone, coloured in red, arises after chemotherapy. In model two many subclones are present in the primary tumour that can vanish, survive or even expand after treatment.

(Adapted from L. Ding et al.[91])



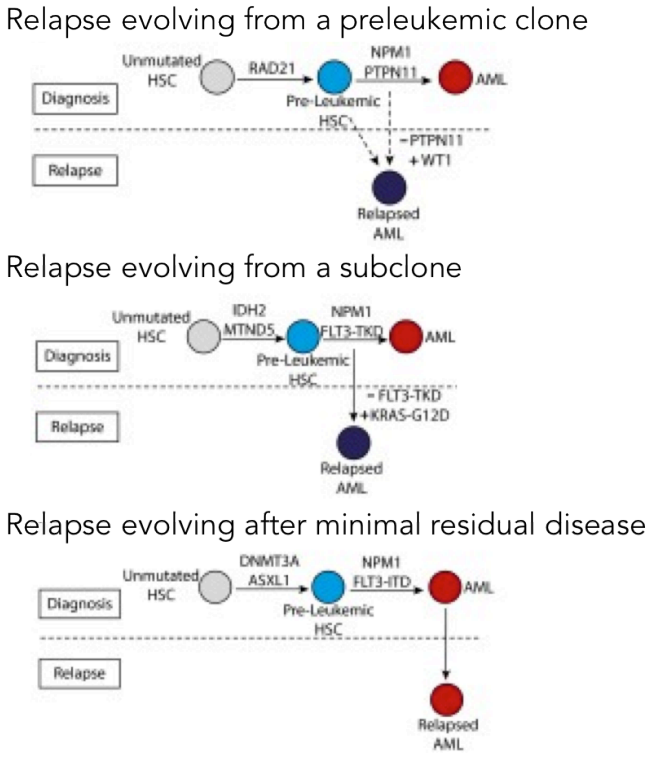
The authors highlighted a significantly higher rate of transversions in the relapse compared to primary mutations. Compared to transitions, transversions are a kind of mutations more difficult to arise in the genome spontaneously and have been

associated to mutagenic factors such as tobacco smoke[92] and cytotoxic agents.[93] Thus, the relapse specific mutations identified may be actively induced by the chemotherapeutic treatment. However, an alternative hypothesis could be that the low coverage they used to perform WGS (25X) is insufficient to identify low frequency mutations at exordium.

In seven out of the eight patients the primary tumours contained mutations associated to preleukemic state of the leukaemia like DNMT3A, IDH1, IDH2, PHF6 and RUNX1. These "landscaping" mutations were also found in some remission samples by R. Corces-Zimmerman et al.[61] In particular, this study shows that the preleukemic HSCs persisted in remission. Using targeted amplicon sequencing to detect the mutations at exordium and relapse in their three patients, the authors highlighted three different scenarios (Figure 1.18): in one patient the relapse probably originated from a preleukemic clone that persisted after chemotherapy and gained additional activating mutations leading to relapse formation; in the second patient a resistant subclone gained new mutations; in the third patient the relapse contained the same mutations identified in primary leukaemia, suggesting that the treatment was not completely effective.

Figure 1.18: three patients exhibit different patterns of evolution from the primary tumour to the relapse leukaemias. In the first patient, the relapse evolved from a preleukemic cell that acquired new proliferating mutations. In the second patient, the tumour arose from a subclone (like model 2 in Figure n.). In the third patient, the relapse leukaemia is the genetically identical to the primary AML, unveiling the survival of leukemic cells also at remission (MRD).

(Adapted from Corces-Zimmerman et al.[61])



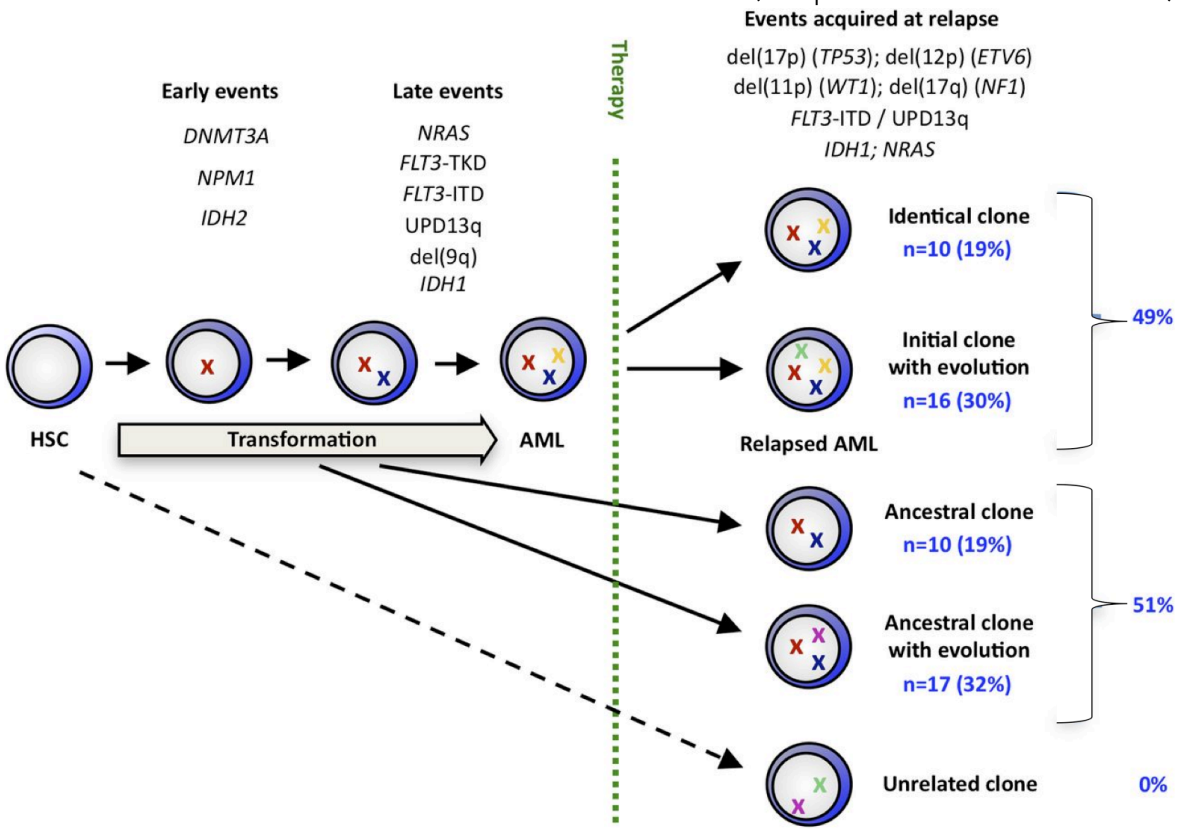
Krönke et al.[94] in 2013 used, instead, SNP array techniques to analyse a cohort of 53 adults with NPM1 mutated AML. They were able to observe all 4 possible combinations of the evolutionary behaviours described by the previous studies mentioned above: half of the NPM1 mutated AMLs appear to evolve from the dominant clone of the primary leukaemia, half from an ancestral clone containing only early "landscaping" mutations and without late events (Figure 1.19); in both cases, they observe the clone with or without the acquisition of new mutations. Similarly to the two studies previously described, new mutations are found in the evolution of the dominant clone or subclones, but, at the same time, they disclose the occurrence of relapse without acquisition of additional mutations in

the primary tumour clones. The mutations acquired at relapse were landscaping (e.g IDH1) and activating (e.g. NRAS). They never observed relapse originating from clones unrelated to the primary leukaemia.

Interestingly, they uncovered the presence of an inverse correlation between the number of mutations shared by the primary and relapse disease and the time to relapse, suggesting that more time is needed for relapse when additional evolution is needed to restart the disease.

Figure 1.19: Evolution of relapse in a cohort of 53 AML patients with mutated NPM1. Below is depicted the summary of the observations made by Krönke et al. in their study on the evolution of relapse. They observed 4 main categories of evolution: the major AML clone is present at relapse with or without evolution or an ancestral clone (that lacks late events from the primary clone) is present at relapse with or without additional mutations. Early events, in general, consist of mutations that affect so-called "landscaping" genes while late events consists of activating mutations. Mutations acquired at relapse can be of both types.

(Adapted from J. Krönke et al.[94])



We think that these three studies glimpse the mechanism of AML relapse emergence that still needs to be further elucidated. Indeed, they highlight some

patterns in AML evolution but, at the same time, have some weak points that need to be solved. Ding et al.[95] use WGS for their analysis, which gives a very broad look at the genome, but at low coverage, we believe that, at higher sequencing depth, some of the relapse specific mutations might be found also in the primary tumour. On the other hand, Corces-Zimmerman et al.[61] and Kronke et al.[94] use microarray for their studies: this technology gives results at higher definition but restricts the analysis exclusively to mutations already identified and inserted in the array *a priori*; furthermore, the former study had a very little cohort, only 3 patients, and the latter examined only a subset of all AMLs, AML harbouring NPM1 mutations. Indeed the combination of the strong points of each study described would give a more thorough representation of the relapsing AML genomic features.

2. Aim of the project

After first remission, about three out of four AML patients relapse within 5 years from the original diagnosis. Relapse can be the result of persistence of leukemic clones or subclones and effective molecular markers would aid both in the identification of patients more prone to relapse and the determination of remission. The aim of our study is, therefore, to combine the strong points of all the studies described in paragraph 1.8 and to identify, through next generation sequencing (NGS), a group of mutations or a signature that might allow the prediction of the treatment outcome at the exordium of AML and to test the hypothesis that the chemoresistant phenotype is characterized by mutations that enable the cells to survive the pharmacological treatment and expand in the secondary tumour. Our experimental plan includes the whole exome analysis of 30 pairs of primary/relapsed AML samples using NGS, to identify relapse-specific mutations, the clonal-evolution bioinformatics analysis to determine the evolution scenario that best fits the observed data and the identification of patterns of mutations or pathways that correlate with the relapsing disease.

3. Materials and methods

In this section we describe all the methods we tested and effectively used for our analysis. We present multiple datasets because these analyses have been performed through the years while the data collection advanced. Paragraphs from 3.2 to 3.6 are dedicated to methods refinement. Lastly, in Paragraph 3.7, we describe the cohort of samples and the pipeline of analysis effectively used for our project and in paragraph 3.8 we define the list of AML driver genes we used to highlight the possible players for tumour formation.

3.1 The dataset

The human AML and APL samples were collected at the University of Bologna, University of Rome Tor Vergata, University of Torino and University of Udine (Italy), and genomic DNA was isolated using standard protocols. In order to identify tumour-specific mutations, we compared each leukemic sample (bone marrow) to the corresponding normal DNA, isolated from blood cells at the time of clinical and molecular remission of the disease. Exome-capture was performed using the SureSelectXT Human All Exon v.1, v.2, v.4 and v.5 (Agilent Technologies) following the manufacturer's specifications. For 25 samples a different exome capture was used for the analysis: for the primary and remission samples of BO6, BO7, BO9, BO10, BO11, BO13, BO14, BO16, BO21, BO22, BO23, BO24, BO27, BO28 TrueSeq rapid capture kit (Illumina) was used for

exome enrichment; on the contrary for primary and remission samples of BO5, BO8, BO12, BO15, BO17, BO18, BO19, BO20, BO25, BO26 and BO29 the capture has been performed with NextEra (Illumina). Whole-exome sequencing was performed with the Illumina Hiseq 2000 platform with 101 bp paired-end reads for all the patients except for UD14 that was sequenced on the Illumina NextSeq machine. General characteristics of the patients are reported in the Results section (Paragraph 4.2.1).

3.1.1 Dataset for alignment testing

Patients analysed in this context were 3 APLs and 2 AMLs collected at the University of Rome Tor Vergata. WES was performed through Illumina GAIIx with 76 bp paired-end reads after DNA isolation with standard protocols; for capture we used SureSelectXT Human All Exon capture kit v.1 and v.2, respectively for AMLs and APLs.

3.1.2 Cohort of samples for mutation calling

Human AML samples used for this analysis are described in Table 3.1. We didn't use hAPL#Mi1 and hAPL#Mi4 samples because of their low quality. Alignment to the reference genomes (hg19) was performed using the Burrows-Wheeler Aligner (BWA)[96].

Table 3.1: Mutation calling cohort – patient characteristics.

Sample ID	Cytogenetic analysis	FAB classification	NPM1	FLT3	Age at Diagnosis
AMLp6	NK	NA	wt	wt	45
AMLp7	NK	NA	mut	wt	53
hAML#Mi3	NK	NA	mut	wt	71
BO1	NK	M1	wt	wt	32
BO2	+8, t(2;10)(q33;p13)	M5	wt	wt	42
BO3	NK	M1	mut	mut	34
hAML#Mi7	inv(9)(p11q13)	M5	mut	mut	50
TO1	NK	M5	wt	wt	67
TO2	NK	M5	mut	mut	73
TO3	NK	M1	wt	wt	58
UD1	del(6); del(11)	M0	wt	wt	33
APLp2	t(15;17)	M3	NA	NA	61
APLp3	t(15;17)	M3	NA	wt	42
hAPL#Mi6	t(15;17)	M3	NA	mut	54
hAPL#Mi7	t(15;17)	M3	NA	mut	56
hAPL#Mi8	t(15;17)	M3	NA	NA	24
hAPL#Mi9	t(15;17)	M3	NA	NA	38
hAPL#Mi10	t(15;17)	M3	NA	mut	68
hAPL#Mi11	t(15;17)	M3	NA	NA	43
sAML#Mi1	t(15;17)	M3	NA	wt	24

3.1.3 The "Bologna cohort"

The cohort of coupled samples used to test the performances of CNV calling algorithms was composed of 23 leukaemia patients. For these patients the group of Prof. G. Martinelli at the University of Bologna sequenced (WES) the primary tumour and the normal sample (except for one case) and at the same time performed SNP array analysis on the tumour sample. In Table 3.2 are reported the quality of the SNP array platform, the presence/absence of WES sequencing of tumour and normal samples and the platform used for the SNP array analysis. We decided to exclude some patients from the analysis: BO9 and BO10 because they were analysed on a different SNP array platform, BO12 because the normal was not sequenced by WES; BO14, BO16 and BO22 because they did not pass the quality filters.

Table 3.2: Summary of the sequencing data available for the samples from the Bologna cohort. For every patient we report the quality of the SNP array platform, the eventual WES sequencing for the tumour (TUM) or the normal (NORM) samples and the platform used for the SNP array analysis.

quality	sample name	TUM	NORM	SNP array
ok	BO9	OK	OK	GenomeWideSNP_6
ok	BO10	OK	OK	GenomeWideSNP_6
ok	BO11	OK	OK	CytoScanHD
ok	BO12	OK	missing	CytoScanHD
ok	BO6	OK	OK	CytoScanHD
ok	BO5	OK	OK	CytoScanHD
ok	BO13	OK	OK	CytoScanHD
not passed	BO14	OK	OK	CytoScanHD
ok	BO15	OK	OK	CytoScanHD
not passed	BO16	OK	OK	CytoScanHD
ok	BO8	OK	OK	CytoScanHD
ok	BO18	OK	OK	CytoScanHD
ok	BO19	OK	OK	CytoScanHD
ok	BO20	OK	OK	CytoScanHD
ok	BO21	OK	OK	CytoScanHD
not passed	BO22	OK	OK	CytoScanHD
ok	BO23	OK	OK	CytoScanHD
ok	BO24	OK	OK	CytoScanHD
ok	BO25	OK	OK	CytoScanHD
ok	BO26	OK	OK	CytoScanHD
ok	BO27	OK	OK	CytoScanHD
ok	BO28	OK	OK	CytoScanHD
ok	BO29	OK	OK	CytoScanHD

3.2 Comparing mappers of the sequencing reads to the genome

Alignment of the short reads produced by the NGS technology consists in the mapping on the human reference genome (in our case hg19/Grch37) of each read, defining its exact genomic coordinates. This task is performed comparing each base belonging to the read with the bases of the reference genome in order to detect the region with the highest similarities. This step is tricky in Bioinformatics because it is time consuming and memory consuming (the genome is $3X10^9$ bp long and the sequencing platforms usually produce hundreds of

millions reads for each sample). We compared the performances of two methods in the alignment of short reads to the genome, approaching the problem from diverse perspectives: BWA[96] and Novoalign[97]. BWA uses the Burrows-Wheeler transformed (BWT)[98] algorithm. Thanks to the BWT indexing technique, it is very efficient and allows the presence of mismatches and gaps. Novoalign[97] uses the Needleman-Wunsch[99] global alignment with gap penalties, an old method that the authors were able to adapt for alignment optimization.

In order to compare the output of the two methods, we used SEAL[100], a comparative tool primarily designed to evaluate short read aligners. Given a set of parameters (sequencing error, indels, coverage, …), SEAL is able to simulate the sequencing data, producing fastq files that contain reads which resemble an NGS output. The reference genome can be either already existing or created by the program and the reads are generated choosing positions on that reference genome from a uniform distribution and fragment sizes (for paired end experiments) from a normal, in order to simulate the characteristics of real reads. Afterwards, the chosen alignment methods (in our case BWA and Novoalign) are run on the simulated fastq files in order to permit the comparison of the results obtained through alignment with the simulated genome of origin.

3.2.1 Pre-processing for alignment testing

Reads were filtered along machine quality assessment keeping only reads marked as 'N' by the machine (N: Not failing the quality filter, with a quality score

threshold of 2), the alignment was performed in parallel with BWA and Novoalign with standard parameters. After alignment, we applied the GATK pipeline[101] including realignment, duplicates identification (Picard)[102] and quality recalibration. Mutations were called using MuTect[103] and annotated with ANNOVAR.[104]

3.3 Comparing mutation calling algorithms in WES-AML samples

We decided to measure the differences among MuTect[103] and SomaticSniper[105] because the former is widely used in WES analysis in the literature and the latter was primarily used for the definition of single nucleotide variants in AML[44] resulting in two distinct landscapes on the same cohort of leukaemia patients. The approach of the two methods is rather analogous: given the reads observations at that site, MuTect calculates for each not-reference site the likelihood for the tumour to carry or not that variant, SomaticSniper calculates the probability of all the possible genotypes. The main divergence between the two methods locates in the reads filtering procedure for the tumour and control samples (see Results section 4.1.2). For both algorithms we functionally annotated SNVs using the MutationAssessor database[106] and filtered out synonymous variants or those falling in non-coding regions.

3.3.1 Mutation calling with SomaticSniper

We followed the pipeline guidelines for SomaticSniper[105], removing duplicated reads before mutation calling with the MarkDuplicates software present in Picard version 1.68[102]. Only variants with mapping quality and somatic quality values over 40 were included in the successive analysis.

3.3.2 Mutation calling with MuTect

WES data have been pre-processed according to GATK best practices[101,107] through local realignment, duplicate marking and base quality recalibration. We identified SNVs in our samples using MuTect version 1.1.4.5 with the initial tumour Log Odd Discovery set to 6.8 (calculated on the basis of the expected number of mutations per Megabase in the TCGA 2013 publication[44]). An additional filter was applied to the output SNVs on the minimum read depth: a minimum of 8 reads should be present in the normal samples and 14 in the tumours, as suggested by the authors.

3.3.3 Validation

We validated high-frequency mutations (≥ 25%) through Sanger sequencing, after PCR amplification with custom primer pairs for each mutation. All PCR products were evaluated on a 2% agarose gel, then sequenced in both directions with Big Dye Terminator reactions and loaded on an ABI PRISM 3730xl DNA analyser. To analyse the sequences we used the Sequencing Analysis 5.2 software.

For low frequency variants, Sanger sequencing was ineffective because it has low sensitivity and is unable to validate variants present in less than ~25% of the reads (our experimental data). For these reasons, we validated variants with VAFs <25% using the IonTorrent platform, choosing the regions in order to explore all the possible groups of frequencies and, when possible, selecting both variants detected only by MuTect or by SomaticSniper and variants detected in common by both algorithms. Ion Torrent sequencing data were aligned to the reference genome (hg19) with BWA. For each SNV, we counted the number of reads carrying the variant in the tumour and in the remission. We considered validated the variants with a p-value <0.001, calculated with the one-sided test on equality of proportions in the comparison between the number of reads carrying the variant in the normal or the tumour samples (with a maximum number of reads in the normal supporting the variant allele ≤3).

3.4 Comparing CNVs detection methods

Copy number variants are usually easier to be detected in WGS data than in WES. In fact, WES data are characterized by the discontinuity of the regions tested and are affected by the capture of the target regions, which is susceptible to diverse affinity of the probes. Nonetheless, many tools have been developed for the detection of CNVs from NGS data and we decided to compare some of the most used methods for CNV identification in WES samples:

- Cn.mops[108]: the acronym stands for Copy Number estimation by a Mixture

Of PoissonS. In order to uncouple the copy number detection from the inner variability of exome coverage across the genes, this tool models the coverage at each genomic position. Through a Bayesian approach it decouples the real signal and the noise and uses Poisson mixture models to detect noise and reduce false discovery rates (FDR);

- CONTRA[109]: the COpy Number Targeted Resequencing Analysis tool uses the ratio of coverage depth in tumour and control samples to detect regions that vary in the number of copies. To overcome the general bias produced by this approach, CONTRA: i) normalizes the coverage depth in the two samples; ii) uses the log-ratios at the base level to circumvent errors due to GC-content; iii) corrects for imbalanced library size; iv) estimates the log-ratio variation binning the regions and using interpolation;
- ExomeCNV[110]: uses depth of coverage and alternative allele frequencies to detect CN regions. The ratio of coverage depth in the tumour and normal samples is normalized for the total number of reads in each sample and adjusted to have a total median of 1. CNVs are, then, recognized through hypothesis testing: when the Poisson distribution of one region's coverages can not be associated to a normal distribution, a CNV can explain this deviation. A ρ parameter is used to distinguish between deletions and amplifications;
- Control-FREEC[111]: this tool is similar to the abovementioned ExomeCNV, but it adds finer normalization steps, adjusting for GC-content and

handling also contaminated control samples;

- VarScan2[112]: uses Fisher's exact test to extract the regions where the log2 ratio of depth of coverage in the two samples undergoes a significant change. It, then, uses circular binary segmentation (CBS, part of DNAcopy R package) to identify the exact regions that have alterations in the number of copies.

In particular, to call copy number variant regions from WES data in our cohort of patients, we used the Control-FREEC tool, adjusting for contamination and using a minimum read count threshold of 50 reads. The window used to compare regions was set to 50'000 and the target region was restricted to the portion of the genome covered by the capture design of the Sure Select kit version 5 (Agilent) plus the complete sequence of exons covered only partially by that regions (without UTRs).

3.4.1 SNP array analysis

Aroma[113] suite (CRMAv2 R package) was run with default parameters and GLAD segmentation; for our patients only the leukemic sample was analysed and, because we lacked the SNP array of remission samples, a summary of all the samples was used as normal reference.

Nexus[114] suite, provided by BioDiscovery, used as control sample HapMap counts.

3.5 Gillespie's stochastic simulation algorithm

It is of paramount value the possibility to describe the behaviour of a system over time, given defined constraints. Stochastic time evolution equations that describe a finite system are unmanageable analytically (except for very simple cases) and also numerical solutions can be difficult. Simulations allow studying the complex dynamics of finite states that evolve in time designing variable trajectories. They start from a set of given probabilities for variables, and randomly change; simulations can be repeated many times in order to observe the distribution of results.

The Gillespie's algorithm[115] is capable of simulating numerically the evolution in time of a model in an exact manner. In origin, it was designed for chemical systems but it can be extended to all phenomena that can be represented with reactions following the laws of mass action. Here, we present the description of the procedure that directly solves the time simulation described in paragraph 4.1.5.1 of the Results. Given a population i=N at time *t*, the number of individuals in state x is $X_i(t)$. At each time point of the simulation, we calculate the vector $\boldsymbol{X}(t) \equiv (\boldsymbol{X}_1(t), \dots, \boldsymbol{X}_N(t))$. Starting from time 0 when $X_{t_0} = x_0$, the system changes over time on the basis of a set of reactions ($R_j, \text{with } j = 1, \dots, M$) that introduces birth, death, changes of state (…) in the existing population. At each step, the system is updated following the state change vector (v_j, that describes all the possible population changes introduced by the reactions in the system) and the propensity function ($a_j(x)$, which is the probability of occurrence of a reaction in

the infinitesimal time step $[t, t + \partial t)$) with values sampled at that particular step.

At each step, the direct method substitutes t with the successive time step $t + \tau$ and x with $x + v_j$. τ is defined as follows: r_1 and r_2 are sampled such that $\tau = \frac{1}{a_0(x)} \ln\left(\frac{1}{r_1}\right)$ with $a_0(x) = \sum a_j(x)$ and $j = \min\left(\sum_{i=1}^{j} a_i(x) > r_2 a_0(x)\right)$. In this way, the more the population or the reaction rate increases, the more the time-step decreases to better describe what is happening in the system.

In our study we reproduced the evolution of the tumours and their relapses in order to know *a priori* the subgroups forming the tumour populations in the two cases and to be able to retrace the steps that led to that result. The R package used to run the Gillespie's method in order to obtain model solutions is GillespieSSA[116]; we used the direct method "D" to get the exact Monte Carlo procedure. The seed has been sampled at every cycle in the interval [-1'000'000, 1'000'000] and is reported in the output.

3.6 Clonal analysis methods

For every solution identified with the Gillespie's algorithm for the tumour and relapse couples, we built a dataset reporting the position, the base change, the number of reference and alternative reads in the three samples (primary tumour, relapse and the normal) and their relative copy number. Further description of the model and how the dataset information was extracted is reported in the Results section, paragraph 4.1.5.1. On the datasets produced by the Gillespie's algorithm, we were able to test four clonal reconstruction methods taken from the

most important papers published in the field:

- Clomial[70]: is the compression of "Clonal decomposition using binomial models". This algorithm is able to handle multiple samples at the same time in order to increase the statistical power. To estimate clonal genotypes and their frequencies uses an expectation maximization algorithm. We used the Clomial R package with random seed set to 1. We ran Clomial with a suggested number of clones from 1 to 5, the best model was chosen as the one with the lowest value of BIC (Bayesian Information Criterion, used to select the model with the highest likelihood). We decided to limit our research in the range [1,5] because the amount of time needed by the algorithm grew significantly with the number of expected clones given in input. Furthermore, in the majority of the examples, the best models resulted to be those with 2 expected clones;
- Expands[69]: is the acronym for "Expanding Ploidy and Allele Frequency on nested Subpopulations". This algorithm models cellular frequencies as probability distributions and, then, uses hierarchical clustering to group mutations with similar probabilities. In this way the mutations with similar behaviour (in the probability distribution) should be grouped also if they have different frequencies, as for the CNV regions. We used the Expands R package with random seed set to 11. We used the parameters suggested by the authors and, in particular, we set the number of amplicons per mutated cell to 50, the precision to 0.018 and the minimum cell frequency to 0.01. For clustering, we restricted the dataset to those variants having

finite cell frequency distributions;

- PyClone[117]: uses a hierarchical Bayesian statistical model to estimate cellular prevalence together with allelic imbalance introduced by CNVs. The group of mutations and the number of groups in the populations are computed simultaneously through Bayesian non-parametric clustering. Interestingly, it uses beta binomial emission densities stating that it overcomes simple binomials for high variability in the prevalence of mutations. Since allelic prevalence and number of copies are strictly connected to each other it also uses flexible prior probabilities for the genotypes. Beta binomial density distribution was the statistics chosen to run PyClone for 10'000 Monte Carlo Markov Chain iterations. Alpha and beta parameters for Dirichelet Process were set to 1. Tumour content was set to 1, because, for the majority of samples in our analysis, we do not have the information about purity. Expected sequencing error rate was set to 0.005, which is the upper bound for the error rate in our analyses. All the other parameters were set to default values;
- SciClone[118]: uses Bayesian mixture modelling of beta distributions to cluster the mutations by VAFs similarity. It uses only variants falling in copy number neutral regions of the genome. We used sciClone R package with random seed set to 11. We used the default parameters: the minimum depth of the region was set to 60, the maximum number of clusters to 10 and the copy number margins to 0.25.

All the pieces of information retrieved at the precedent steps were necessary to

perform a good analysis of the clones in our cohort of patients. For every patient we ran PyClone, analysing the triplet of the samples all together. In particular, the random seed was set to 7 and the burning region to 1000, as suggested by the authors.

3.7 The pipeline defined for our analysis

In the first part of the results we surveyed a group of methods to define the ones that would better apply to our specific cohort of relapsing AMLs. The pipeline used to preprocess the sequencing data is reported in Figure 3.1. Starting from the FASTQ given in output by the Illumina platform, we firstly filtered the reads that passed the quality test, performed inline during the sequencing. We retained only the reads marked with 'N' (those not failing the quality filter with a threshold of 2) by the sequencer. We, then, aligned all the reads to the human genome (hg19) with BWA mem, the latest version enhanced for time consumption and precision. We did not filter out unmapped reads, we, instead, marked (but not removed) the duplicates with Picard MarkDuplicates version 1.84, using lenient validation stringency. The reads were, then, realigned to the genome to improve the precision around indels. The realignment was performed through GATK version 2.8, with stringency set to lenient, using as known databases for the indels 1000 genomes phase 1 and Mills and 1000 genome gold standard. The base qualities were recalibrated, in order to have more reliable values, using the recalibration tool from GATK (version 2.8) with dbSNP137 as gold standard. This

pipeline was applied to every sample following GATK guidelines. However, we integrated this pipeline in a Python script in order to automatize the process and reduce the computational time for the analysis. Once all the triplets of samples completed this first part of the pipeline, they were jointly used for the search of variants.

Figure 3.1: Schematic representation of the analysis pipeline.

3.7.1 Mutation calling with MuTect

Identification of the single nucleotide variants (SNVs) was performed running MuTect (version 1.0.27783), with reference databases COSMIC version 54 and dbSNP version 135. We performed the MuTect analysis 6 times for every patient in the following combinations of comparisons: tumour *versus* remission, relapse *versus* tumour, relapse *versus* remission and the *vice versa* comparisons. The MuTect results were filtered retaining only mutations marked as KEEP and removing both not covered positions (UNCOVERED) and SNPs (DBSNP) for every comparison. For each patient we, then, built a unique file containing all the positions identified as mutated at least once. For each mutated position we used a Python script to restart from the bam files and count again the reads carrying or not the variant in all the three samples. Afterwards, an R script was used to label

every mutation, as described in the Results section (paragraph 4.1.3), on the basis of their presence in the primary, the remission and/or the relapse samples after recounting.

3.7.2 Calling of Indels with Pindel

Pindel was run on the following couples of tumours: primary *versus* remission, relapse *versus* primary and relapse *versus* remission. We filtered only regions with coverage greater or equal to 10, more than 3 alternative reads and BWA quality score greater or equal to 20. Annotation was performed with ANNOVAR.

3.7.3 Cleaning the contaminated samples

As described in paragraph 4.2.2 of the Results, four relapse samples in our cohort (UD1, UD4, UD11, UD12) were contaminated with the DNA of the donor for the heterologous bone marrow transplant. Likely, we were able to retrieve the DNA of the donors. Therefore, for the contaminated relapse samples we needed to subtract the SNVs, indels and CNVs detected in the donors DNA. For this reason, we sequenced for WES analysis the donor DNA and ran the abovementioned pipelines to identify the variants. We, then, eliminated all the SNPs and the indels identified in the donors from the relapse sample of each corresponding recipient patient. For the CNVs we, instead, removed from the results the regions overlapping for more than 10% with the donor CNVs. Though harsh, this action was necessary to produce reliable results in the successive analysis.

3.8 A list of AML driver genes

The list of AML driver genes, reported in Table 3.3, was built considering the output of many driver-calling tools from the literature: MutSig[45], MuSiC[47], OncodriveFM[119], TUSON[120], DOTS-Finder[46] and "Vogelstein" (a list of driver genes reported in Vogelstein et al. 2013 paper [121]). All those methods were run on a set of 200 AMLs and all the genes identified as drivers by at least one of those methods were retained in our dataset.

The functional categories for AML drivers were derived from the TCGA paper[44] and covered the 90% of the drivers in our list described above.

Table 3.3: The list of driver genes used for our analysis.

Gene Symbol	MutSig	MuSiC	Oncodrive	TUSON	DOTS_Finder	Vogelstein	PipelineCalls
ASXL1	x	x	x	x	x	x	6
BCL2				x		x	2
BCOR					x	x	2
CALR					x		1
CBFB					x		1
CBX7					x		1
CEBPA	x	x	x	x	x	x	6
CREBBP				x		x	2
DNMT3A	x	x	x	x	x	x	6
EZH2	x	x		x	x	x	5
FLT3	x	x	x		x	x	5
HNRNPK					x		1
ID3				x			1
IDH1	x	x	x		x	x	5
IDH2	x	x	x		x	x	5
KIT	x	x	x			x	4
KMT2D				x		x	2
KRAS	x	x				x	3
MEF2B				x			1
MEF2BNB-MEF2B				x			1
MIR142		x					1
MXRA5	x						1
MYC				x			1
MYD88				x		x	2
NPM1	x	x	x	x	x	x	6
NRAS	x	x	x		x	x	5
PAPD5	x						1
PDSS2	x						1
PHF6	x	x	x		x	x	5
PTPN11	x	x	x			x	4
RAD21	x	x			x		3
RUNX1	x	x	x	x	x	x	6
SF3B2				x		x	2
SMC1A	x	x					2
SMC3	x	x					2
SRSF2	x					x	2
STAG2	x	x			x	x	4
TET2	x	x	x		x	x	5
TP53	x	x	x	x	x	x	6
U2AF1	x	x	x		x	x	5
WT1	x	x	x	x	x	x	6

4. Results

The main aim of this thesis is the investigation of the evolution of the mutational landscape of AMLs from the exordium to the relapse in order to identify possible drug resistant lesions that might allow the prediction of patient's outcome if they are present in the primary disease. We characterised of a dataset composed by 30 patients for whom we collected the leukemic blasts of the primary tumour and the relapse tumour and the blood cells at the remission phase of the disease. We sequenced the whole exome of all the samples and then compared the triplets of samples to uncover possible commonalities and differences that would explain the underling evolution of the disease. In particular, we divide the Results section in two main parts. In the former, we assess the performances of different available bioinformatics methods and choose the best methods for our purposes. In the latter we show the results obtained on our cohort of patient with the selected methods.

4.1 Selection and refinement of methods to improve WES data interpretation

Undeniably, there is an expanding quantity of NGS analysis tools for all the steps that need to be executed to interpret exome sequencing data (e.g. alignment, point mutation mutation callers). These tools exploit slightly different needs and have different performances. The first task of our bioinformatics endeavour is,

therefore, to compare and select the best methods able to portray our cohort of patients.

4.1.1 BWA is more suitable than Novoalign for mapping reads in our cohort of leukaemia patients

Burrows-Wheeler Aligner (BWA)[122] is, nowadays, the most commonly used method to map sequencing reads to the reference genome in WES studies and it is the recommended mapper in the GATK guidelines for DNA high throughput sequencing.[107] Even though this method is generally recognized to have high performances, it is possible that, on specific datasets, other methods may perform better. We had the possibility to test Novoalign[97], which is a mapper under license and has been used by other groups in our field[123]. We compared the performances of these two methods using two approaches:

- SEAL[124]: a tool that allows to test alignment methods, creating an in silico dataset with parameters set by the user;
- we analysed a in house set of WES sequencing of APL and AML samples (not validated) in order to unveil the overlap of the results obtained with the two methods and eventually highlight whether one of the two was more comprehensive than the other.

SEAL produces fastq files containing reads with specific characteristics chosen by the user (see Methods section 3.2); in particular our custom set had a "genome" length of 38'000'000 bp, we selected this length because it corresponded to the

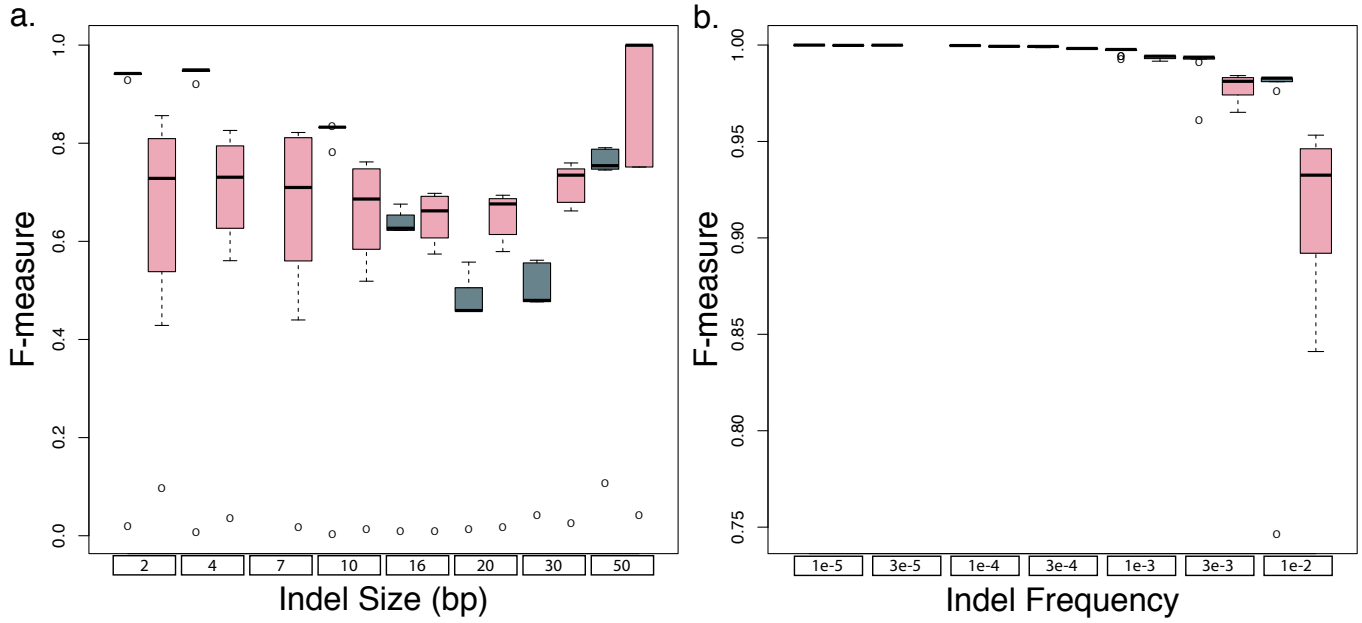
exons covered by the SureSelect kit version 1, the fragment-length was of 76bp and the fragment-count of 36'000'000 bp. The program only allows changing the genome indels size and indels frequency and we decided to evaluate the changes in alignment performance modifying each condition separately. The results of our analysis are reported in Figure 4.1. We used the F-measure as a parameter for the quality of alignment, it was calculated as follows:

$$F - MEASURE = 2 * \frac{(PRECISION * RECALL)}{(PRECISION + RECALL)}$$

where Precision is the rate of true positives over all the positive calls $\left(\frac{TP}{TP+FP}\right)$ and Recall is the rate of true positives over all the real positives $\left(\frac{TP}{TP+FN}\right)$: higher results correspond to better performances.

Varying the size of the insertions or deletions, we observed that BWA performs better than Novoalign when the indels are small (< 10 bp), while Novoalign have higher F-measure when the indels are bigger than 16 bp. In contrast, the frequency of the indels has a minor impact on the goodness of the alignment and both methods perform very well for frequencies lower than 3×10^{-3}.

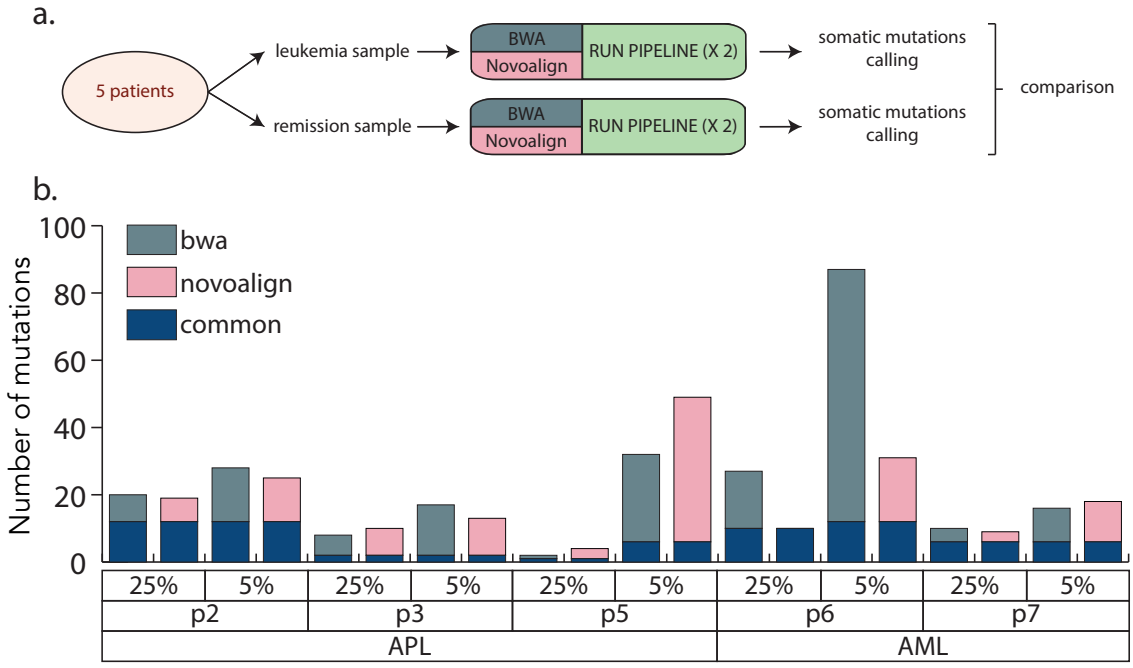
Figure 4.1: BWA and Novoalign performances differ on an in silico dataset. We maintained fixed the indels frequency and vary their size (a) or, alternatively, maintained fixed the indels length and vary their frequency (b). We simulated the alignment to the reference genome with BWA (green boxes) or Novoalign (pink boxes) and evaluated the correctness (F-measure) of the mapping with SEAL.



We, then, decided to test the two methods on a real dataset (pipeline depicted in Figure 4.2.a). For this purpose, we used coupled tumour and remission WES samples from 5 leukaemia patients: three APLs and two AMLs (for a description of the dataset see Materials and Methods section). For every patient, both samples were mapped to the reference genome using BWA or Novoalign. Afterwards, following the downstream analysis suggested by GATK, mutations were called using the MuTect pipeline, comparing remission and tumour samples aligned with the same mapper. As shown in Figure 4.2.b, we observed that, in the majority of cases, the mutations identified after alignment with the two independent methods are not highly concordant and the overlap of the results is often lower than 50%. For each of the two alignment algorithms, the private mutations identified (i.e. the mutations identified specifically only by one method)

are many and BWA seems to identify better low frequency mutations. Only a vast validation of the identified mutations would really allow determining the effective performances of the two aligners; however, we performed a validation analysis of the mutations called only after alignment with BWA. The validation rate we obtained is high; therefore, we know that, in our hands, BWA works well (see results described in detail in the next paragraph).

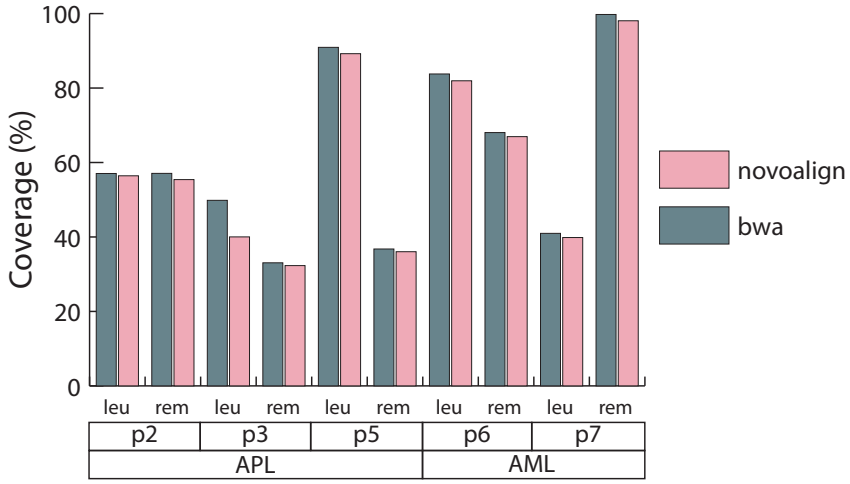
Figure 4.2: Comparison of BWA and Novoalign on a real dataset of AMLs. a. Workflow used for this comparison. b. Each bar indicates the number of mutations identified in common by both programs (dark blue) and, alternatively, only by one of the two (grey: BWA; pink: Novoalign). The results of the analysis are shown both for mutations identified at high frequency (VAF > 25%) and at low frequency (VAF > 5%). P2, p3, p5, p6 and p7 are the patients used for the analysis: the first three were diagnosed with Acute Promyelocytic Leukaemia (APL) and the last two with Acute Myelogenous Leukaemia (AML). For every patient we grouped together mutations at high frequency (having a VAF over 25%) and all mutations (having a VAF higher than 5%).



Another useful and important parameter to take into consideration when evaluating an alignment tool is the coverage of the mapped reads within the targeted region of interest. Since, in order to perform WES enrichment, we used the SureSelect Human All Exome V1 kit (Agilent Technologies), we were able to

calculate this parameter, in terms of percentage of the targeted region covered by the mapped reads in our sequencing dataset for AML patients. The performances are reported in Figure 4.3: despite BWA has always a slightly better coverage than Novoalign, the two methods are nearly comparable and there is no striking difference.

Figure 4.3: Percentage of the coverage on target of the reads mapped with BWA and Novoalign on our 5 leukaemia patients. The coverage is reported as percentage of the total amount of bases covered by the SureSelect Human All Exon kit V1 (Agilent technologies). P2, p3, p5, p6 and p7 are the patients analysed in this circumstance: the first three were Acute Promyelocytic Leukaemia (APL) patients; the last two were Acute Myelogenous Leukaemia patients (AML). For every patient we report the coverage of the samples collected at exordium of the leukaemia (leu) and remission (rem).



We compared the performances of BWA and Novoalign considering three capacities required to define a good aligner:

- The capacity to manage small indels: for frequencies of 1 x 10^{-3} and lower, the two methods are almost comparable and very good in the alignment task. On the contrary, varying the indels size, the two tools show opposed capabilities: BWA is significantly better than Novoalign at mapping reads when the indels size is less than or equal to 10 bp; they perform similarly when the indels size is around 16 bp and Novoalign operates better on

indels between 20 and 50 bp of size. In our project, we want to study the presence in our patients only of small indels and we decided not to search for structural variants (i.e. indels of big size), for which the methods developed up to now are actually still unsatisfactory. Based on these observations BWA is more suitable in the context of small indels alignment;

- The SNVs called after alignment: the two methods result in about half common mutations and half aligner-private mutations; in particular, BWA seems to manage more efficiently low frequency mutations. We did not expect a similar discrepancy of identified variants originating by the used aligner. However, since, in our hands, mutations identified at the bottom of the pipeline using BWA as mapper, have high validation rates, we think that it is a good mapper for our study;
- The coverage of the target regions: this parameter was comparable for the two tools and BWA gave slightly better results than Novoalign.

For all the argumentations mentioned above and because it was not under licence, we decided to use BWA for all our further analyses.

4.1.2 MuTect allows mutation calling at low Variant Allele Frequency

Accurate mutation calling in WES of tumour-control couples of samples is still a challenge in cancer studies, because there are many sources of error that may arise both during the preparation (PCR amplification, enrichment capture), the

sequencing of the samples (machine error) and the bioinformatics analysis of the mutations (mapping to the reference genome, correct quality assessment). Moreover, intrinsic peculiarities of some tumours, in particular blood tumours, such as the heterogeneity of the population and the contamination by the tumour of the normal sample, used as control, pose an additional layer of complexity. Growing evidence of the high variability in the mutations called through different algorithms[125] has encouraged the scientific community, firstly, to try to develop new outstanding tools for mutations detection[126] and, secondly, to build methods that combine the results of multiple tools in order to obtain more complete and reliable results.

4.1.2.1 Two analysis pipelines strongly disagree in mutation calling over a set of leukaemia patients

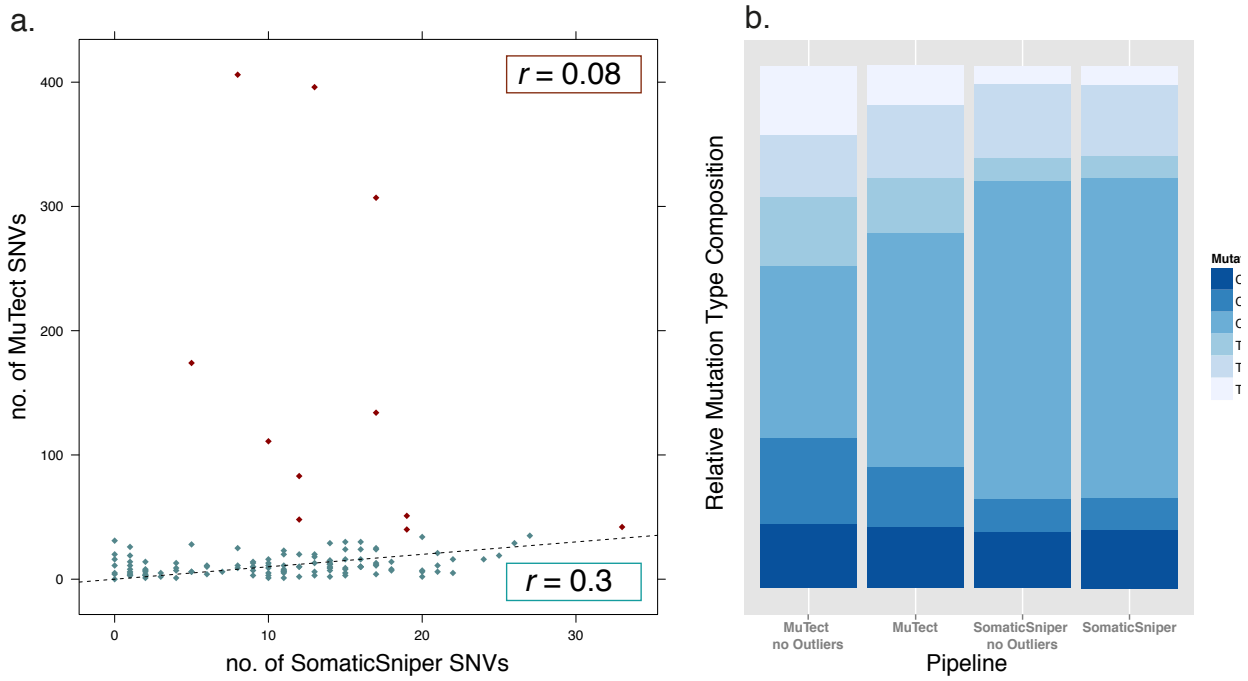
We noticed that two successive studies[44,45] on the same cohort of patients gave very discrepant results (Figure 4.4). We used the supplementary information of the two papers in order to collect the number of mutations and the type of substitutions (i.e. the detected base change) identified for each patient. We uncovered a substantial difference in the variants identified by two different pipelines. The first study[44], performed at the Washington University, used SomaticSniper as variant caller; the second study[45], a pan-cancer analysis made at the Broad Institute, used MuTect. The two studies report very different numbers of mutations from the WES analysis of the same 133 AML patients (Fig 4.4.a), with an average of 11 *vs* 24 mutations (0.03-1.7 *per* Mb vs 0-13.53 *per* Mb) *per*

patient, respectively. Particularly different was the maximum number of mutations identified in single patients: 33 vs 406, with seven patients showing a very high number of mutations detected by MuTect, never highlighted before in AML. We called these patients "hypermutated" because they are conform to the hypermutation definition[127] having a number of somatic mutations in the coding regions dramatically above the median mutation rate of the other AMLs.

Correlation coefficients calculated for the number of mutations *per* patient, including or excluding these outliers, are 0.08 and 0.3, respectively, underlining the little similarity in the results obtained by the two analyses. Furthermore, also the type of mutation changes, identified in the two studies (Fig 4.4.b), shows two different genomic landscapes for the same pathology and the same cohort of patients.

Figure 4.4: Two mutation callers give significantly different results on a public dataset. a. Number of mutations identified for each patient by SomaticSniper on the x axis and MuTect on the y axis. The red dots identify patients with a very high number of mutations when analysed with MuTect (outliers). Pearson's correlation coefficients were calculated including (red square) and excluding (green square) these patient outliers. The black dashed trend line indicates the expected number of identified mutations assuming that the analysis with the 2 pipelines gives exactly the same number of mutations. b. Percentage of each type of mutations listed (Mutation-Type) on the total number of mutations identified by each pipeline. Also in this case the analysis was performed including or excluding outlier patients (no Outliers).

(Adapted from Bodini et al.[128])



We, then, selected all the datasets that were analysed with the same preprocessing pipeline and were available from the TCGA data portal and re-analysed 131 AML WES datasets using MuTect. Considering the union of the SNVs identified with the new MuTect analysis and the previous set of SomaticSniper mutations, we characterised a new and more comprehensive AML mutational landscape. Indeed, analysing the TCGA bam files with MuTect, we could detect the presence of many more SNVs in the AML tumours than the ones identified by the TCGA project. In particular, we identified 9150 putative false negative SNVs and the majority of them (5983) with VAFs <10%. In addition, we

scored in multiple patients the presence of recurrent SNVs in the same position, never described before in leukaemia patients, for the following genes: RNF2, TP53BP2 and RASA1 (Tab 4.1). Again, also the SNVs identified in these genes were usually present at VAFs <10%. These genes encode for proteins involved in regulation of cell proliferation or differentiation. Moreover, using DOTS-finder[46], a tool developed in our laboratory aiming at the identification of driver genes from a list of patient's mutations, they were classified as putative drivers. These results underscore that the genes we found mutated at low frequency may play a relevant role in cancer progression.

Table 4.1: Putative driver genes identified by two pipelines, found recurrently mutated in the AML cohort analysed. The table shows the percentage of patients carrying a mutation in each listed gene as identified by one (TCGA) or both pipelines (TCGA+MuTect). The genes reported have been pinpointed by DOTS-Finder as putative driver.

	Percentage of mutated patients			Percentage of mutated patients	
Hugo gene symbol	TCGA + MuTect	TCGA	Hugo gene symbol	TCGA + MuTect	TCGA
DNMT3A	29%	29%	MYH11	2%	0%
FLT3	12%	10%	NACA	2%	0%
IDH2	11%	11%	NUP214	2%	0%
TP53	11%	11%	PHF6	2%	2%
KMT2C	11%	1%	PRDM1	2%	0%
IDH1	10%	8%	THRAP3	2%	2%
NRAS	10%	8%	TPR	2%	0%
RUNX1	8%	6%	TSC1	2%	0%
KRAS	7%	5%	ABL2	2%	0%
ATP2B3	5%	2%	ASXL1	2%	1%
EP300	5%	0%	ATM	2%	0%
KIT	5%	5%	ATP1A1	2%	0%
MED12	5%	1%	BCL11A	2%	0%
NF1	5%	1%	BRCA2	2%	0%
PTPN11	5%	5%	BUB1B	2%	0%
HIP1	5%	0%	CACNA1D	2%	0%
TET2	5%	5%	CD74	2%	2%
U2AF1	5%	4%	CLTCL1	2%	0%
AKAP9	4%	0%	EBF1	2%	0%
ALK	4%	0%	ERG	2%	0%
CREBBP	4%	0%	EZH2	2%	2%
KDR	4%	2%	FANCA	2%	0%
ROS1	4%	0%	FUBP1	2%	0%
SF3B1	4%	1%	GAS7	2%	1%
TCF12	4%	0%	GMPS	2%	0%
WIF1	4%	0%	GPHN	2%	0%
APC	3%	0%	KDM5A	2%	0%
ELF4	3%	1%	KIF5B	2%	0%
GATA2	3%	2%	LCP1	2%	0%
KMT2D	3%	0%	MAX	2%	0%
MSN	3%	0%	MLLT4	2%	1%
NCOA2	3%	0%	MYB	2%	1%
NOTCH2	3%	1%	MYC	2%	1%
PBRM1	3%	0%	MYD88	2%	0%
SETBP1	3%	2%	MYH9	2%	0%
TRRAP	3%	0%	NFE2L2	2%	0%
WT1	3%	2%	NTRK3	2%	2%

ARID1A	2%	1%	NUMA1	2%	0%
BCR	2%	0%	NUP98	2%	0%
BRD4	2%	0%	PAX3	2%	1%
CHEK2	2%	0%	PAX5	2%	0%
CIC	2%	1%	POU5F1	2%	1%
CNOT3	2%	0%	RAP1GDS1	2%	1%
CSF3R	2%	1%	SETD2	2%	0%
EGFR	2%	2%	SLC34A2	2%	0%
ETV6	2%	1%	SRSF2	2%	0%
EXT1	2%	0%	STAG2	2%	2%
FANCD2	2%	0%	STAT5B	2%	0%
JAK2	2%	1%	SYK	2%	0%
KDM6A	2%	2%	TRIP11	2%	0%
LPP	2%	0%	TSHR	2%	0%
MSH6	2%	1%	USP6	2%	1%

As shown in Table 4.1, for all genes harbouring non silent SNVs and present in the Cancer Gene Census database, we could not score differences in the frequency of mutations in our patients cohort for well known AML drivers (e.g. DNMT3A, IDH1, IDH2, FLT3, RUNX1, TP53, NRAS). However, combining the results of the two pipelines, we identified 58 novel recurrent SNVs in genes present in the Cancer Gene Census database, even if they were present in 5% of the patients at most. However, though novel mutations were infrequent among patients, 43% of the patients carried at least one of them, pinpointing their relevance in this dataset.

Furthermore, we analysed the new results using the Fisher's exact test in order to define which genes are mutated in a mutually exclusive fashion (mutually exclusive genes) and which genes are mutated simultaneously in the same patients (co-occurring genes). If we do not correct for multiple hypothesis testing (as in Ley et al.[44]), we can recapitulate the same results of the TCGA publication[44]

for mutual exclusivity. In contrast, we can find many more co-occurring genes (542 vs 12, considering only genes and not gene functional cathegories). However, if we correct for multiple hypothesis testing, almost no genes show significant co-occurrence or mutual exclusivity. In agreement with our observations, a subsequent publication by the same authors described an identical scenario[64]. Indeed, applying a multiple hypothesis testing correction the authors found only two couples of co-occurring genes. We think that these results are given to the low frequency of the new mutations we identified in these patients. Likely, the analysis of a larger cohort would deliver more reliable results for testing co-occurrence and mutual-exclusivity.

Furthermore, as stated above we identified a very high number of mutations in 7 patients. For 3 cases this is very likely due to the presence of non silent SNVs in mismatch repair genes: MSH6 in two hypermutated patients and PMS1 for the third patient. In contrast, there are no evident or easy explanations for the remaining 4 cases. However, we have to underline that our analysis is restricted to SNVs, so we cannot exclude the presence of other types of mutations, such as indels or structural variants in mismatch repair genes also for the remaining patients.

4.1.2.2 Testing and validating the two mutation calling pipelines on a cohort of 20 leukemic patients

Considering all the discrepancies that were highlighted by our comparison in the results obtained with the two pipelines, we decided to try to investigate further

the reasons underlying these discrepancies. Therefore, we performed a thorough validation of the mutations identified by both methods in order to determine if one of the two had a greater error rate and, possibly, the rationale for such errors. To this aim, we analysed by WES a cohort of 20 leukemic patients containing 19 primary AMLs (8 of which APLs) and one secondary APL. In agreement with the data we reported using the TCGA dataset, also on our cohort of patients, the SNV landscapes identified by the 2 pipelines were largely different: SomaticSniper revealed a total of 194 SNVs and 178 mutated genes, versus 463 SNVs and 412 mutated genes revealed by MuTect. The commonalities amounted for 161 SNVs and 150 mutated genes. Indeed, the frequencies of SNVs per patient are different: 9.7 in SomaticSniper vs 23.2 for MuTect (0-0.6 per Mb vs 0.1-4.37 per Mb). Moreover, similarly to the paper mentioned above[45], MuTect characterized one hypermutated patient: the secondary APL, harbouring 131 mutations. In this patient we did not detect mutations in mismatch repair genes, but the hypermutated phenotype could have been induced by the external chemotherapeutic agents used for treatment.

The genes that were previously known to be cancer genes (as reported by The Cancer gene census, CGC[129,130]) identified by the two pipelines were 14 for SomaticSniper and 21 for MuTect (Table 4.2), underling again the impact of the discrepancies found and the importance of using a good mutation caller.

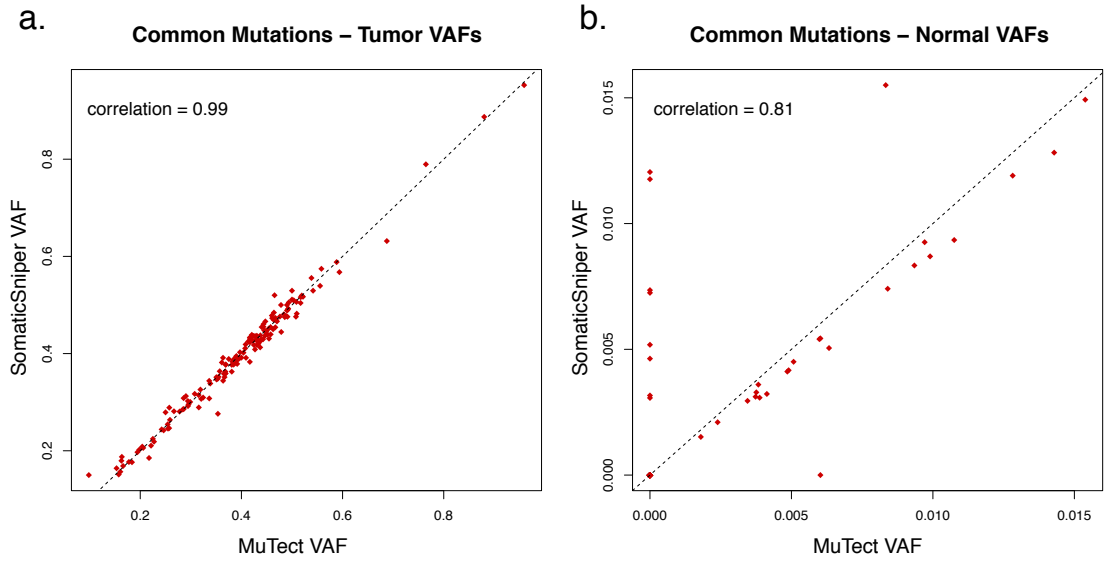
Table 4.2: Number of mutations identified by SomaticSniper and MuTect in our cohort of 20 leukaemias. For each patient (sample), we report the number of SNVs uncovered by SomatiSniper, by MuTect, by both pipelines (Common). Moreover, for each group we report the genes already known to play a role in cancer development that are part of the Cancer Gene Census (CGC).

Sample	SomaticSniper	MuTect	Common	CGC genes only in SomaticSniper	CGC genes only in MuTect	CGC genes in Common
AMLp6	0	3	0			
AMLp7	8	23	7			IDH1
hAML#Mi3	7	17	5			IDH2
BO1	13	17	8	NRAS	KRAS	
BO2	8	10	7			IDH1
BO3	13	13	5	DNMT3A		IDH1
hAML#Mi7	7	7	6			FLT3, IDH1
TO1	16	20	15			JAK2, PHF6, RUNX1
TO2	11	17	10			DNMT3A, FLT3
TO3	15	23	14	TET2	MLL3, NRAS	EZH2, NRAS, RUNX1, TET2
UD1	17	24	16			EZH2
APLp2	14	26	12	IL21R		
APLp3	3	22	3		CHECK2	
hAPL#Mi6	11	21	6		ETV6	
hAPL#Mi7	18	43	16		TRIM33	ARID1A
hAPL#Mi8	3	9	3		FGFR2	KRAS
hAPL#Mi9	5	5	5			
hAPL#Mi10	7	10	7			FLT3
hAPL#Mi11	8	22	7		FLT3, NRAS, WT1	KDM6A
sAPL#Mi1	10	131	9		ERBB2, KIAA1549	

For the variants identified in common, the two algorithms have a high rate of agreement on the VAF of the mutations (Figure 4.5), with high Pearson's correlation coefficient both for the frequencies of the variants identified in the

tumour (0.99) and in the normal sample (0.81). The average difference of frequency assignment was around 1.1% (range 0 - 7%). Indeed, these SNVs were easy to be uncovered having little frequency of reads carrying the variant in the normal (about less than 1.5%) and high frequency in the tumour sample (the smallest around 20%). The normal samples show a slightly lower correlation rate because they are at very low frequencies and one read difference results in bigger discrepancies.

Figure 4.5: Variant Allele Frequency of mutations identified by both algorithms. Common mutations generally have very low VAFs in the normal sample (panel b., ~≤ 1.5%) and higher in the tumour sample (panel a., ~> 20%). The Pearson's correlation coefficient calculated for VAFs in SomaticSniper and MuTect of tumour and normal samples are respectively 0.99 and 0.81, reflecting high similarity of counts. Note that in the two panels there is a different scale in order to magnify the differences of normal samples.



The two pipelines especially differ on their capacity to call SNVs with low VAFs: MuTect in our dataset is able to call mutations with as low as 2% VAF, while the lower limit for detection with SomaticSniper is bound at 10% VAF. For this reason, we grouped mutations based on their frequencies: we considered high frequency mutations (HF) those having a VAF >10% and low frequency mutations (LF) those having a VAF ≤10%. SomaticSniper was able to identify 194 HF SNVs and MuTect

found 245 HF SNVs (9.7 and 12.3 SNVs *per* patient, respectively) (Table 4.3); 161 of these SNVs were common, accounting for around 83% and 66% of the total SNVs identified by each pipeline, respectively. The validation of HF SNVs revealed that both methods are very precise with validation rates higher than 89%; in particular the common mutations had a validation rate of 98% (78/161 tested positions).

The SNVs distinguished by only one of the two methods are 33 for SomaticSniper and 84 for MuTect (Table 4.3). SomaticSniper unique SNVs were verified in 43% of the cases (6/14 positions tested) and MuTect unique HF SNVs were always validated (17 positions tested). These results underline that both pipelines leave a portion of mutations undetected (i.e. the mutations that are unique to the other pipeline). Probably, refinement of the methods, aiming at reducing the detection of false positives, conducted to high rates of false negative mutations.

Furthermore, nearly half of the SNVs identified by MuTect were LF SNVs, never detected by SomaticSniper. The validation rate for these LF SNVs was around 80% (48/60 positions tested). LF mutated genes include also 13 known drivers as reported by CGC[130]: CHEK2, ERBB2, ETV6, FGFR2, FLT3, KIAA1549, KRAS, MLL3, NRAS, PDE4DIP, RUNDC2A, TRIM33 and WT1.

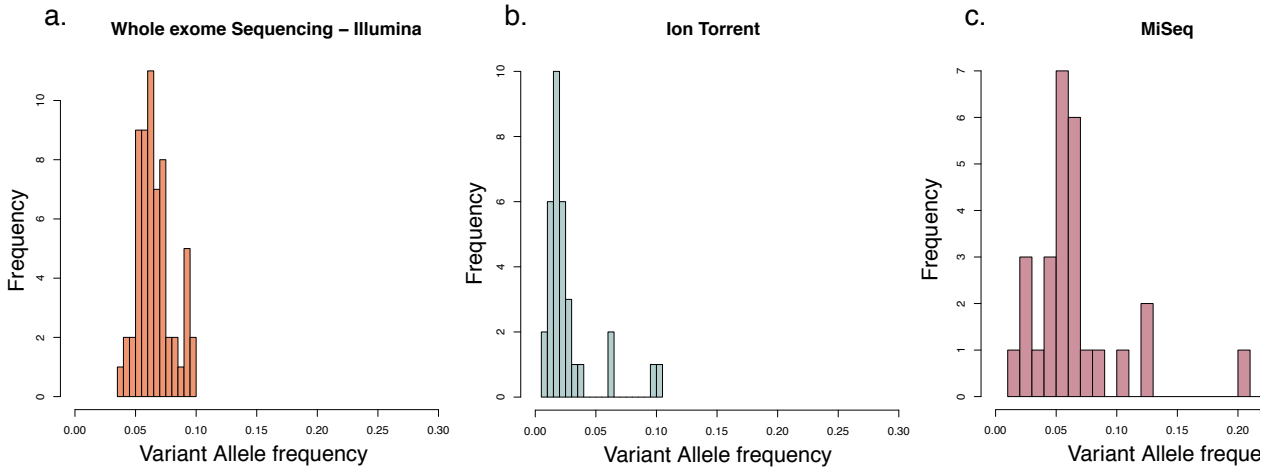
Table 4.3: Mutations identified by the two pipelines and corresponding validation rates. In the table are reported all the mutations identified by the two pipelines (SomaticSniper-ALL and MuTect-ALL) and grouped by identification (COMMON: identified by both pipelines, SomaticSniper-ONLY, MuTect-ONLY). We tested only a portion of all the mutations identified and the numbers are reported for every group. We also separated the results by frequency (higher or equal to 10% and lower than 10%) in order to highlight differences of the methods used.

Frequencies:	All	Tested	Validated	x ≥ 10%	Tested	Validated	x < 10%	Tested	Validated
SomaticSniper-ALL	194	94	84 (89%)	194	94	84 (89%)	0	0	0
MutTect-ALL	463	157	143 (91%)	245	97	95 (98%)	218	60	48 (80%)
COMMON	161	80	78 (98%)	161	80	78 (98%)	0	0	0
SomaticSniper-ONLY	33	14	6 (43%)	33	14	6 (43%)	0	0	0
MuTect-ONLY	302	77	65 (84%)	84	17	17 (100%)	218	60	48 (80%)
Total (SomaticSniper+ MuTect)	496	171	149 (87%)	278	111	101 (91%)	218	60	48 (80%)

LF SNVs identified through WES were tested on two different sequencing platforms, Ion Torrent and MiSeq, to ensure the reliability of the results (see Materials and Methods chapter, paragraph 3.3.3). Comparing Illumina WES frequencies with the output of our validation datasets, we determined that the average distance between VAFs was: for Ion Torrent 6 percentage points (with a minimum of 0.1 and a maximum of 16 percentage points), for MiSeq the average distance was smaller and equal to 3 percentage points but with larger variability (from a minimum of 0.06 to a maximum of 21 percentage points). Compared to the WES data, VAFs resulted decreased in Ion Torrent validation and very similar in MiSeq validation. The validated mutations have frequencies that span from 4% to 10% with a median around 6% in Illumina (both for WES and MiSeq targeted resequencing) and a median of 2% in Ion Torrent. In Figure 4.6, we show the number of mutations we validated, grouped for VAFs in the three experiments. In conclusion, though there is no outstanding correlation, WES seems to

approximate validation frequencies and, indeed, low frequency mutations (<10%, median of 6%) in WES appear at VAFs lower than 10% in IonTorrent (median of 2%) and in MiSeq (median of 6%). We deliberately tested mutations that appeared at more than 4% in WES analysis; both validation technologies confirmed that these were LF mutations, probably at VAFs even lower than estimated through WES. Because the higher the coverage, the more precise can be the VAF determination, with a coverage below 200X (typical of WES analysis), the determination of LF VAFs is restricted to few reads and is more prone to fluctuations.

Figure 4.6: Variant Allele Frequencies of low frequency validated mutations. The three panels show the VAFs of the LF mutations detected respectively through WES (a.), Ion Torrent (b.) and MiSeq (c.). We deliberately tested mutations that appeared at more than 4% in WES analysis; validation technologies confirmed that these were LF mutations, probably at VAF even lower than observed through WES.



In order to ascertain the accuracy of detection of mutations at very low VAFs, we calculated the empirical error rate (i.e. apparent mutation rates at random nucleotide) on our set of sequences obtained with the two validation platforms. Our experimental strategy consisted in defining a consensus sequence for every position covered by more than 10 reads. The reference "correct" base was

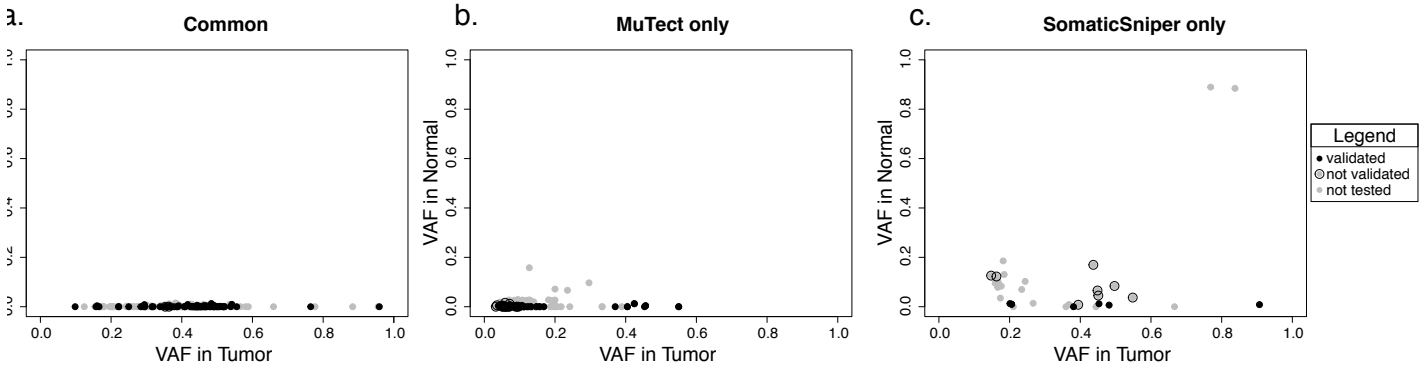
defined as the most called base and reads that differ from the "correct" base were considered as errors. Based on this consensus strategy, we calculated the empirical error rate as the total number of bases that differ from the consensus bases divided by the total number of read bases. The empirical error rate for Ion Torrent was 0.33% and for MiSeq was 0.61% for the control samples (normal) and 0.29% and 0.62% for the tumour samples, respectively. Based on these results, we can be reasonably confident in the trueness of variations appearing at rates above 1%.

Moreover, the possibility of calling an SNV depends on the presence of mutated allele in the tumour sample and its absence in the normal sample: these are the prerequisites of a somatic mutation. Thus, it is possible to find the tumoural variant in the normal sample for technical reasons, caused by mosaicism of the normal or by normal contamination with the tumour sample, in particular when the normal is a remission sample. Pløen et al.[53] described DNMT3A mutations persisting in the normal sample up to 8 years after initial diagnosis (at frequencies up to 50%); those mutations were after represented in the relapse or in a secondary myelodisplastic syndrome revealing the preleukemic nature of the cells harbouring them. As discussed before, the persistence of tumour or preleukemic cells at remission, with frequencies under the detectability level, can eventually cause the relapse in some patients. The evaluation of mutation calling tools in the AML context needs, therefore, to consider this aspect and it is fundamental a certain leniency in the calling test that allows to identify leukemic mutations in the primary also when they are retained in the remission; otherwise the analysis would

loose the most ancient mutations of the leukaemia. Actually, the frequency of variant alleles in normal samples is significantly higher for SNVs unique to one pipeline (Figure 4.7). The average frequency was of 0.001 in the common, 0.004 in the MuTect-ONLY (Welch 2-tailed test, P-value: 0.0004), and 0.1 in the SomaticSniper-ONLY (P-value: 0.005). Notably, common mutations have very low VAFs in the normal, therefore, not validated mutations in SomaticSniper could be the result of a contamination by leukemic cells of the normal sample. SomaticSniper, in fact, was originally designed to call mutations in leukaemias and we noticed that it allows higher VAFs for alternative allele in the normal (Figure 4.7.c). Despite a higher rate of false positives, this leniency allows for the discovery of SNVs that have the "landscaping" characteristics and can result in successive relapse. In our set of patients, we pinpointed mutations with these characteristics in TET2 and DNMT3A that were called "not validated" but necessarily needed to be taken into consideration.

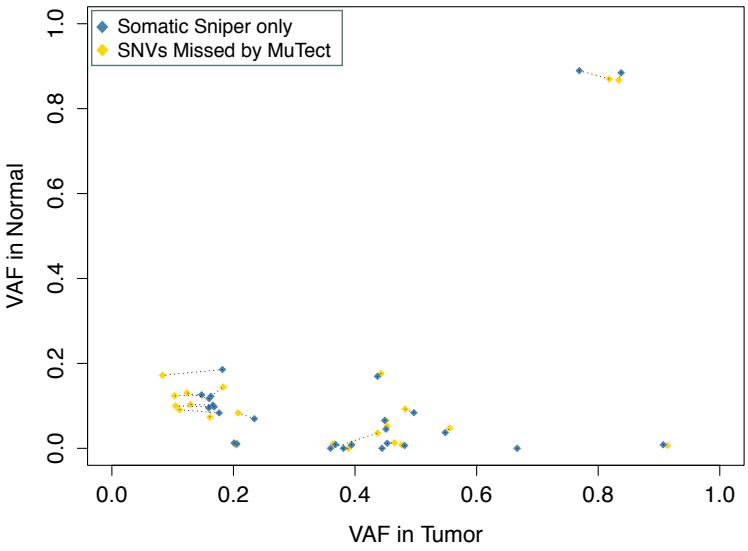
In conclusion, SomaticSniper is more able than MuTect to detect variants present in the normal sample but, at the same time, this characteristic makes it more prone to false positive calls.

Figure 4.7: Normal and Tumour VAFs in called variants. We reported the VAF for tumour (x-axis) and normal (y-axis) for mutations that were commonly identified (panel a) or respectively unique to MuTect (panel b) or SomaticSniper (panel c). Points correspondent to validated mutations are filled in black, not validated mutations are represented with a surrounding black circle.



We retrieved the VAFs reported by MuTect for the SNVs uniquely called by SomaticSniper (*vice versa* was impossible) and we noticed a difference in the frequencies reported by the two algorithms. Indeed, they apply different filters to the reads that result, at the end, in the differences observed in mutation calling. In particular, MuTect applies distinct filters to normal and tumour reads: while the filters applied to the normal are very similar to the filter applied by SomaticSniper, the filters applied to the tumour are stricter in order to increase the possibilities of calling real variants increasing reliability. However, in some cases, we observe VAF counts in contrast to the indicated hypothesis: this can be due to minor differences in filter parameters that in fact are reflected in changes of smaller size (Figure 4.8).

Figure 4.8: Differences in Variant Allele Frequency for the mutations identified only by SomaticSniper. In the graph are reported the VAFs for the tumour (x-axis) and the normal (y-axis) samples for the mutations that were called uniquely by SomaticSniper. Dotted lines connect the points relative to SomaticSniper values (blue) to the points relative to MuTect values (yellow) for the same mutation.



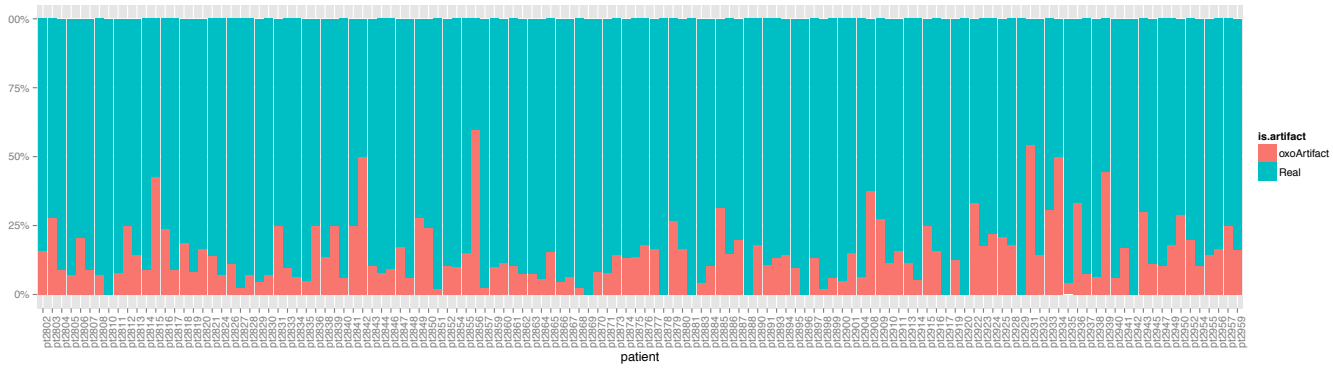
4.1.2.3 The impact of false negatives in the AML data analysis and the choice of a mutation calling method

We compared two mutation-calling methods that have been previously used in the literature to describe AML mutational landscape. The most surprising result was the presence of many false negatives that result in an under evaluation of AML somatic mutations. From the clinical point of view, this can leave concealed patterns of clonal evolution or the presence of prognostic marker that would guide the therapeutic strategies. Indeed, we discovered that a consistent part of the AML mutational landscape has to be uncovered; in particular many mutations at low frequency can be present in leukemic patients that were not characterized with SomaticSniper. It has been demonstrated that oxidative DNA damage during sample preparation can induce artefacts in the mutations[131,132]; therefore we

estimated the possibility for the mutations identified with MuTect to be artefacts calculating the percentage of C>A and G>T transversions on the total mutation number. We verified that some patients present an unexpected high proportion of possible artefact mutations; in particular two patients (pt2856 and pt2931) have more than 50% of C>A and G>T mutations (see Figure 4.9). In these cases a deeper validation should reveal the reliability of the results.

Also the identification of AML hypermutated samples is novel; we still do not know whether the observed mutations can be imputed to technical errors, arise simply by a sample preparation artefact as described above, or are real. Indeed the probability that they are due to oxidative stress is very low since the median percentage of C>A and G>T transversions is 10% (ranging from 7% to 21%). Certainly the presence of hypermutated patients raises many questions on the possible mechanism giving birth to the leukemic phenotype.

Figure 4.9: Percentage of possible artefact mutations in TCGA patients called by MuTect. Each bar in the plot refers to a TCGA patient, on the y axis is reported the percentage of mutations of the C>A and G>T type (red) compared to all the other possible mutations (light blue).



We tested only two of the many somatic mutation callers developed for NGS data analysis; accordingly, our estimate of the number of false negative mutations is likely a limited estimate of the real situation. Nevertheless, the scientific

community already knew this complication and a big effort has been made to find new outstanding method that would outperform the others or that would incorporate the capabilities of multiple tools[126].

Considering all the data obtained and the speculations reported above, we decided to pursue the analysis on our cohort of patients using MuTect, because it allows us to uncover low frequency mutations and have higher validation rates. Although, of course, we must keep in mind the false negatives that also this analysis will bring along.

4.1.3 Calling mutations on triplets of samples

In order to characterize the genetic patterns of relapse formation in AML, we collected, for our cohort of patients, triplets of samples, consisting in exordium, remission and relapse samples. The mutation calling entanglement is even more complex if the analysis is going to be performed on multiple samples collected from the same patient. In this case, calling a mutation in one sample raises the intrinsic probability for the presence of the same mutation in a related sample from the same patient even at very low frequency. Since the purpose of our study is to determine the origin of relapse in AML patients, great relevance is demanded to relapse mutations that were previously existent in primary tumours or remission samples. Therefore, the remission samples in our analysis have not been formally considered simply as normal DNA, and instead we treated them as temporally distinct samples collected during the evolution of the disease.

Practically, we proceeded analysing all the possible couples of samples (tumour *vs* remission, relapse *vs* remission, relapse *vs* tumour and *vice versa*) and calling the variants using MuTect. For all resulting SNVs, we recovered the counts of reads carrying the reference base and the variant base and labelled the variants as:

- Primary tumour specific or Relapse specific when the variant harboured less than two reads or had a VAF lower than 1% (which we consider, as explained in paragraph 4.1.2.2, the lower boundary to distinguish mutations from noise in Illumina HiSeq Sequencing) in the relapse or in the primary tumour sample, respectively;
- Common decreasing or Common increasing if the variant was present in both samples but the difference in the read counts was statistically significant: greater in the primary tumour for the former, greater in the relapse for the latter. We tested the difference in read counts trough two sided Fisher's exact test, using 0.05 as a threshold for significant calls;
- Common Primary – Remission or Common Relapse – Remission in case at least two reads and a VAF of 1% were reached in the remission and one other sample, while the third contained less than two alternative reads;
- Common when the variant was present both in the primary tumour and relapse samples and there was no significant difference in the read counts for the two samples.

This is a simple but effective way to label mutations taking advantage of the joining of multiple samples and looking at the samples as a whole in an

evolutionary context.

4.1.4 Control-FREEC and ExomeCNV outperform other methods in CNV calling

The call of copy number alterations from WES data is to date a challenging issue, mainly because the different affinities for their targets of the probes used for capturing results in coverage fluctuations and because the gaps between the covered regions make even harder the identification of the exact segments with different copy numbers in the tumour cells. Furthermore, it has been recently discovered that not only tumour cells contain copy number alterations but also normal cells, in some cases, present peculiar alterations in the number of copies of genomic regions.

In order to evaluate the performances of different CNV callers from WES data, we took advantage of an additional dataset available in our laboratory. This dataset is only partially overlapping with our main dataset and it is composed by primary tumour and remission samples, which were analysed both by WES and SNP-array analysis. We refer to this dataset as the "Bologna cohort". Of course, in this context, we use the SNP-array analysis as a positive control for the calls of CNVs from WES data, because the sensitivity of this technique, in the regions analysed, is very high and is well established.

Concerning the SNP-array analysis, we disposed of the results of Nexus, the proprietary algorithm for the analysis of CytoscanHD arrays, and we used Nexus'

output as positive control for the evaluation of WES derived CNVs.

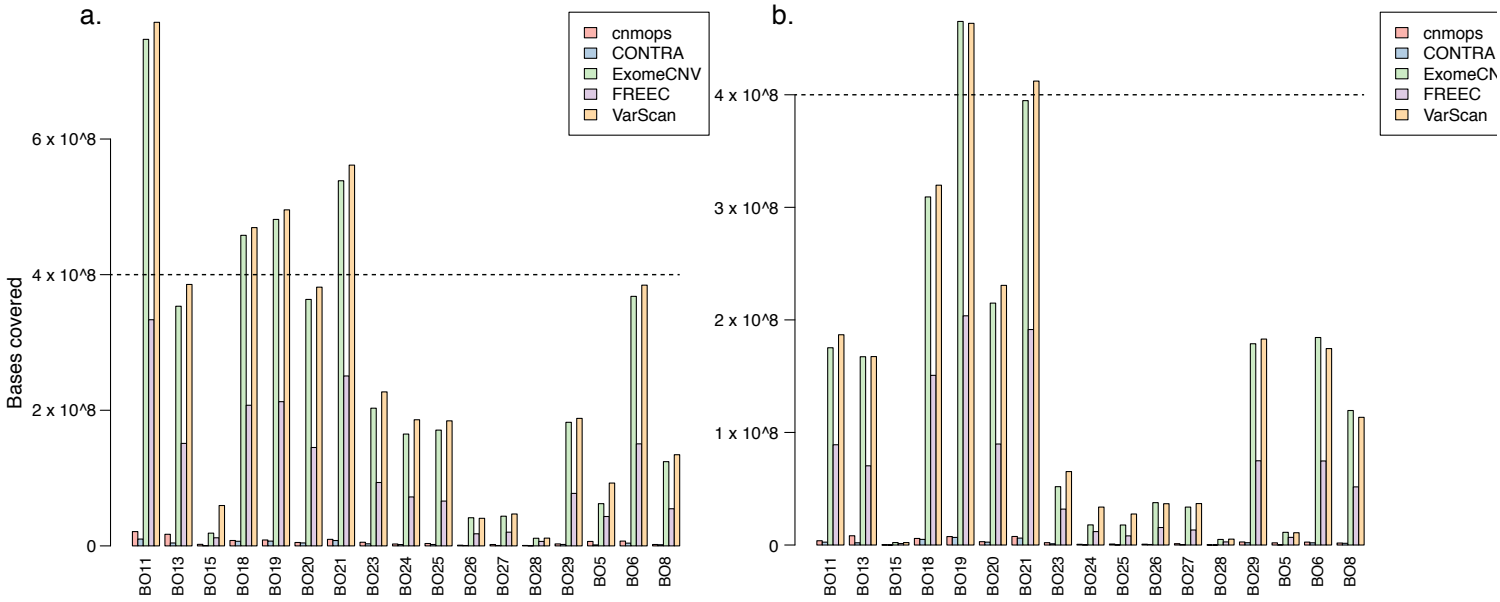
4.1.4.2 Control-FREEC and ExomeCNV have higher accuracies in calling CNVs from WES data

We used five methods to extract CNVs from WES data in order to define which approximates better SNP-array results: cn.mops, CONTRA, ExomeCNV, Control-FREEC, VarScan + DNACopy. For this purpose, we analysed 17 patients from the "Bologna cohort" described in paragraph 3.1.3. For each method, we compared the bed files given in output and containing the calls with the bed files obtained by the Nexus analysis.

First of all, we evaluated if the methods starting from WES data were able to recapitulate the copy numbers variations identified through SNP-array. To address this question, we computed the number of basis of the genome both present in the array output and in the output of each method independently (we called it "coverage", Figure 3.10.a). Nexus output contains only the regions carrying variants in the number of copies (i.e. the number of copies differs from 2), while the tested methods report the information also for copy number neutral regions; as a consequence the "coverage" parameter disclose the extent of the regions containing CNVs effectively targeted by each method (remember that all the methods start from the same reads file). Concerning the coverage, we observed a ranking of performances of the different tested methods: VarScan2 and ExomeCNV gave the best results; Control-FREEC an intermediate performance; cn.mops and CONTRA gave the lowest coverage (Figure 4.10a).

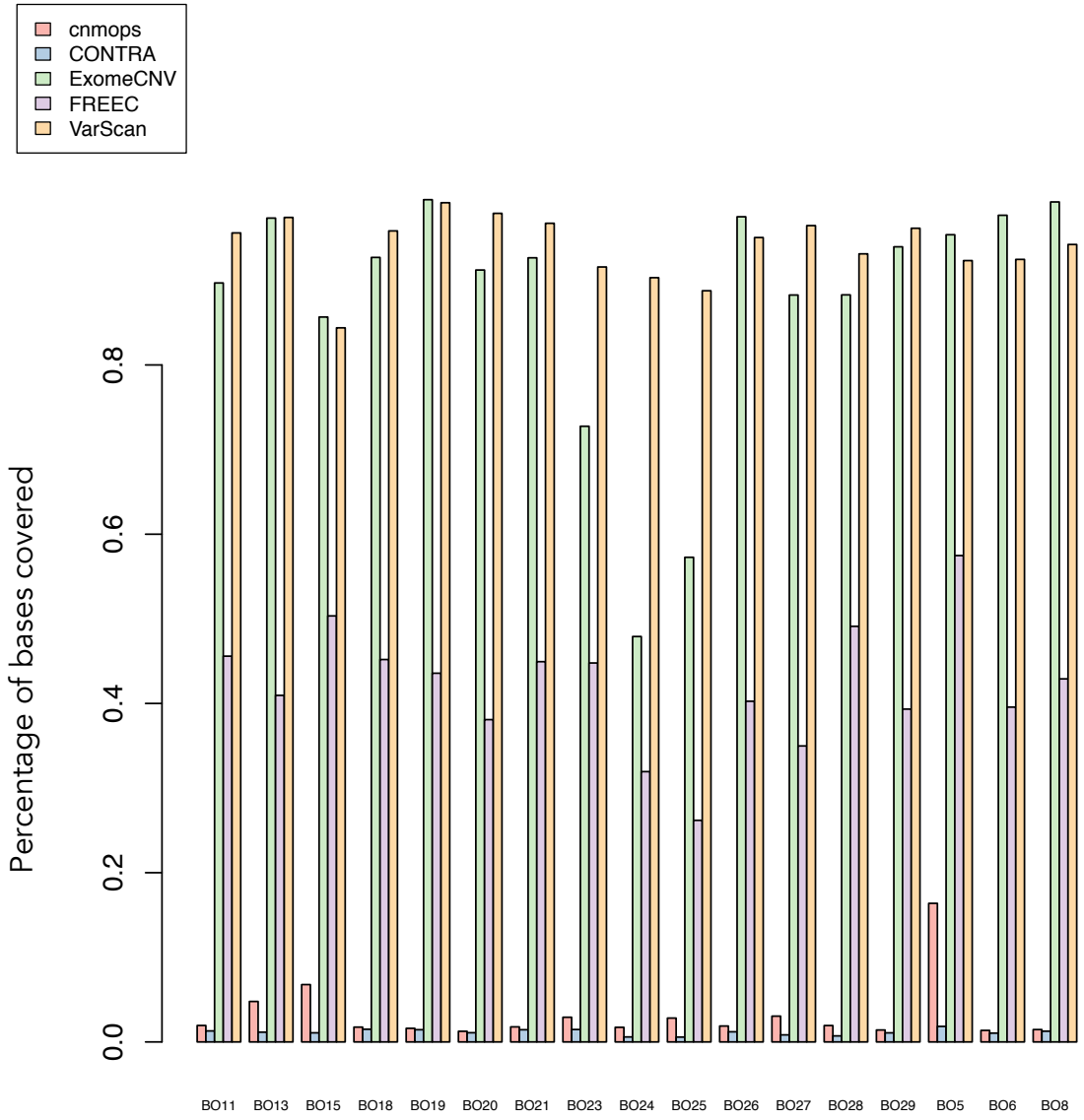
Filtering for exclusively high quality Nexus results, the coverage decreases significantly but not uniformly: some patients have big coverage drops (e.g. BO11), others do not change (BO19), suggesting the absence of association between WES coverage and array quality (Figure 4.10b).

Figure 4.10: Coverage of the overlap of the SNP-array output with the CNV calling obtained from WES data. We calculated the number of bases of the genome present simultaneously in the SNP-array output and in each of the WES-CNV calling pipelines both for raw Nexus results (a) and for quality filter Nexus results (b). Colours of the bars correspond to the methods tested; the dotted line serves as a reference because the two plots have different scales on the y-axis.



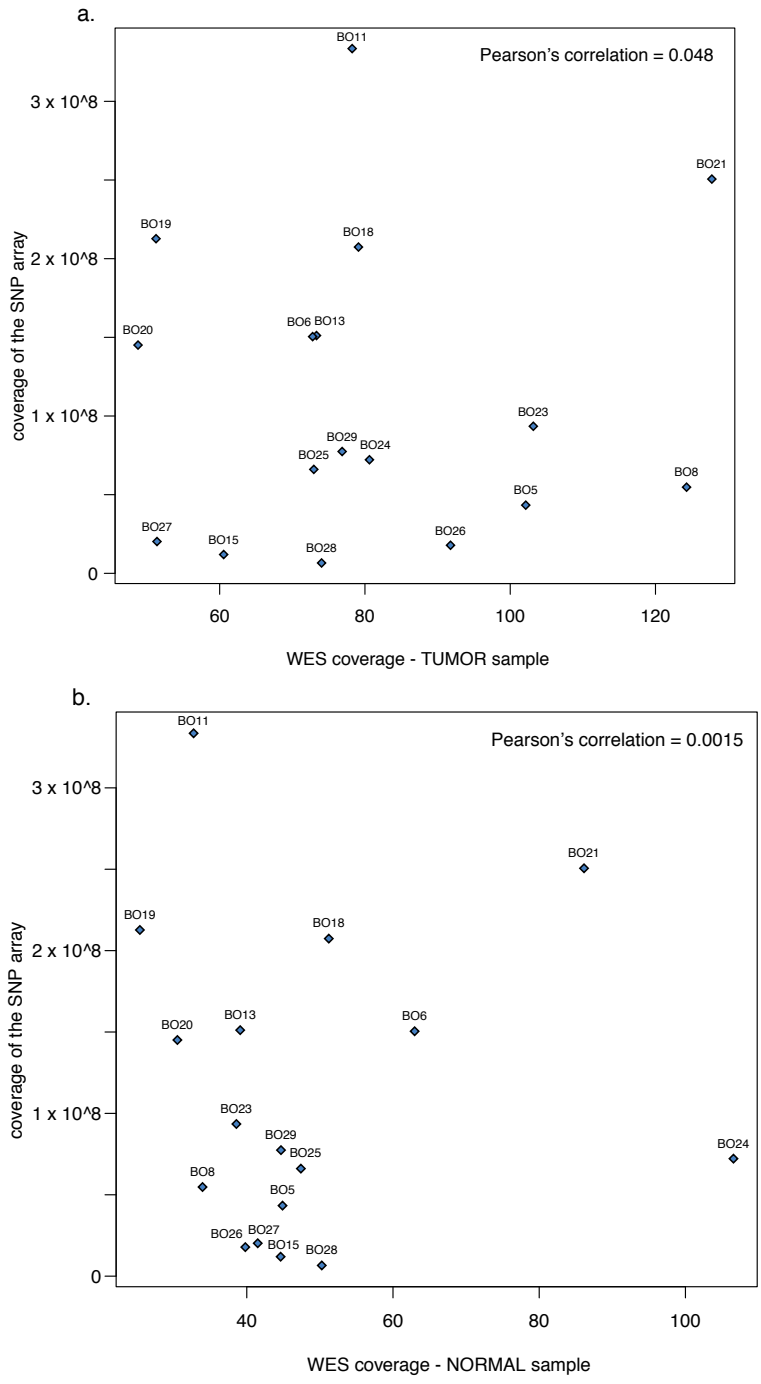
Considering the percentage of basis covered by the SNP array and present in the output of each method instead of the absolute number of covered basis, we recapitulate the same results: CONTRA has the lowest coverage performances, cn.mops is slightly better and variable among patients, Control-FREEC has middle performances and the best results are scored by ExomeCNV and VarScan.

Figure 4.11: Percentage of the overlap regions of the SNP-array output with the CNV calling obtained from WES data. We calculated the percentage of bases of the genome present simultaneously in the SNP-array output retrieved by each of the WES-CNV calling pipelines for raw Nexus results. Colours of the bars correspond to the methods tested.



Indeed, the Pearson's correlation coefficient calculated for WES and array coverage was very low: 0.048 and 0.0015 for tumour and normal samples, respectively, and the distribution of the patients in the plot is sparse (Figure 4.12). In general, for samples with good coverage, the WES coverage does not impact on CNV detection.

Figure 4.12: Correlation between the coverage of the WES regions and the coverage of the SNP-array for tumour (a) and normal (b) samples. Each dot in the plots identifies one patient. The Pearson's correlation coefficient of the distribution is indicated.



In order to compare the CNVs identified by the calling pipelines to the Nexus results, we transformed the absolute copy numbers identified from WES according to these criteria: i) regions with 2 copies were labelled "neutral"; ii) regions with more than 2 copies were labelled "gains"; iii) regions with less than 2

copies were labelled "loss". Afterwards, we were able to produce two confusion matrices for every CNV calling method: one for the GAIN calls and one for the LOSS calls, as described in Table 4.4. However, since Nexus reports only aberrant regions and automatically discards neutral regions, we will lack all the true negatives (TNs) and some false positives (FPs).

Table 4.4: Labels for the construction of the two confusion matrices. Written in grey are the classes missing in our analysis, due to the absence of neutral regions in Nexus output.

Nexus CNV	WES CNV	Gain confusion matrix	Loss confusion matrix
gain	gain	True Positive (TP)	.
loss	loss	.	True Positive (TP)
neutral	neutral	True Negative (TN)	True Negative (TN)
gain	neutral	False Negatives (FN)	.
loss	neutral	.	False Negatives (FN)
neutral	gain	False Positives (FP)	.
neutral	loss	.	False Positives(FP)
gain	loss	False Negatives (FN)	False Positives(FP)
loss	gain	False Positives(FP)	False Negatives (FN)

In this context, to measure the closeness of predicted CNVs identified in WES data to SNP array results, we used accuracy. Accuracy is calculated as:

$$ACCURACY = \frac{(TP + TN)}{(TP + FP + FN + TN)}$$

As already discussed above, the lack of neutral reads in the Nexus outputs allows calculating only partial accuracy. The level of accuracy (reported in Figure 4.13 and summarized in Table 4.5) results very variable among patients: the smallest range for accuracy was obtained with cn.mops on losses going from 0.002 to 0.046 (always presenting very low levels of accuracy), the bigger range was obtained with ExomeCNV for the gains going from 0 to 0.986. The results

obtained from Nexus quality filtered output show higher accuracy variability also for the less performing methods in the detection of copy number losses: CONTRA, the smallest, has an accuracy range going from 0 to 0.22, ExomeCNV in the best case reaches almost 1. Despite, in general, the performances on filtered data do not show very high differences (the median is not strongly affected), the maximum values are significantly improved both for methods showing good CNV prediction capacities (ExomeCNV, Control-FREEC, VarScan2) and for the worse tools (cn.mops, CONTRA). VarScan2 gives unbalanced results: when accuracy is high in CN loss, it is low in CN gain and *viceversa*. On the contrary, ExomeCNV and FREEC seem to be more balanced and they both display good performances. Finally, with our dataset, Cn.mops and CONTRA do not give satisfactory results.

Figure 4.13: Measurements of accuracy of the five WES-CNV calling methods compared to SNP-array results both for LOSS and GAIN CNVs. Using the aforementioned contingency tables, both for gain and loss CNVs, we calculated the accuracy for each method, as shown in the legend, for every patient. The same analysis was performed before and after applying Nexus quality filters to the SNP-array data.

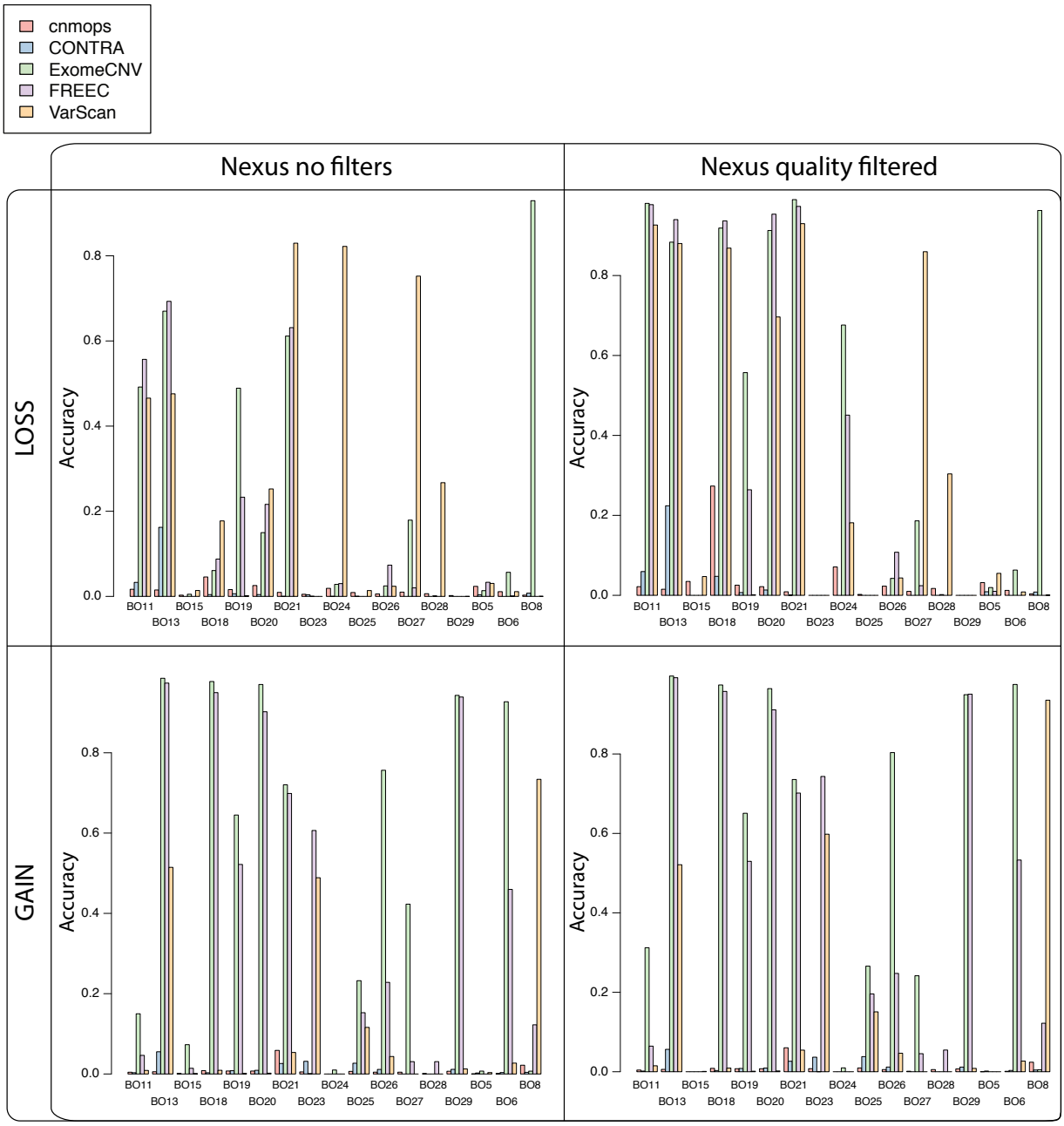


Table 4.5: Summary of accuracy levels identified by the different methods. For every method is reported the median and maximum values obtained on the "Bologna cohort" both for gains and losses. The same set of results was calculated with (Nexus quality filtered) or without (Nexus no filters) filtering the quality of Nexus output.

Nexus no filters				Nexus quality filtered			
Method	Median	Max		Method	Median	Max	
cn.mops	0.005	0.058	gains	cn.mops	0.006	0.06	gains
CONTRA	0.004	0.055		CONTRA	0.004	0.06	
ExomeCNV	0.42	0.995		ExomeCNV	0.31	0.996	
FREEC	0.23	0.97		FREEC	0.248	0.99	
VarScan	0.009	0.73		VarScan	0.009	0.94	
cn.mops	0.01	0.046	losses	cn.mops	0.017	0.27	losses
CONTRA	0.0004	0.162		CONTRA	0	0.22	
ExomeCNV	0.057	0.929		ExomeCNV	0.186	0.99	
FREEC	0.03	0.693		FREEC	0.024	0.98	
VarScan	0.03	0.83		VarScan	0.054	0.93	

Another parameter used to measure the quality of calling is the F-measure, calculated through the following equation:

$$F - MEASURE = 2 * \frac{(PRECISION * RECALL)}{(PRECISION + RECALL)}$$

where Precision is the rate of true positives over all the positive calls $\left(\frac{TP}{TP+FP}\right)$ and Recall is the rate of true positives over all the real positives $\left(\frac{TP}{TP+FN}\right)$: higher results correspond to better performances.

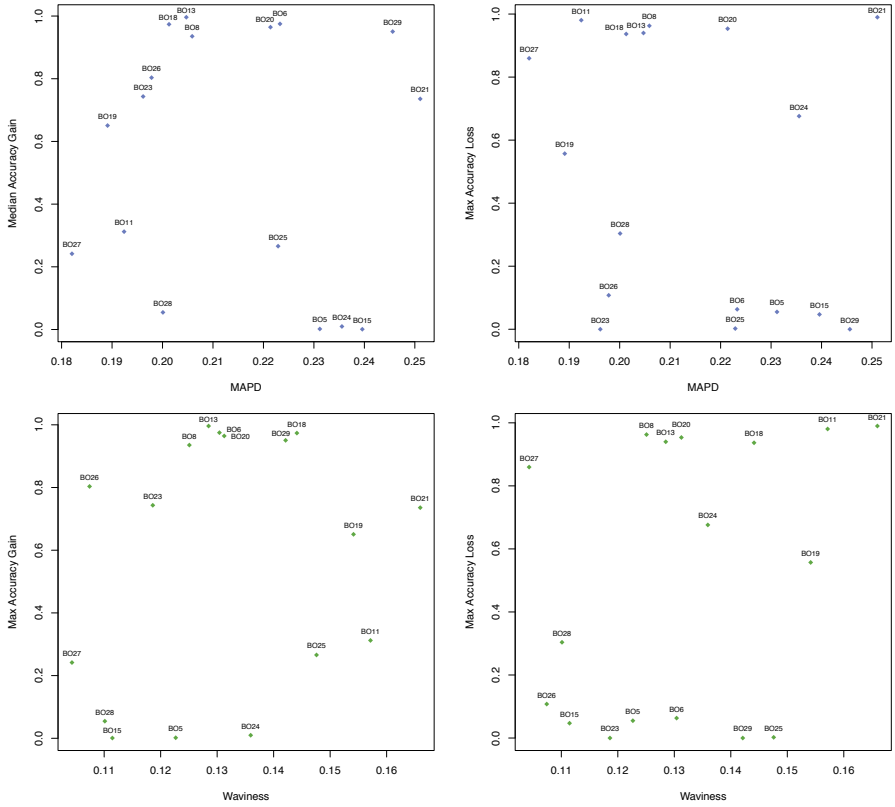
In absence of True Negatives, F-Measure is nearly equivalent to Accuracy and, indeed, as expected in our context, calculating F-measures for our dataset, we obtained results very similar to the one obtained measuring accuracy (data not shown).

Since we observed that the CNVs of some patients resulted problematic (e.g. none method is able to reach 0.1 of accuracy on BO5), we speculated that some intrinsic characteristic of the sample could impair the capacity to identify CNVs in

those samples either biological characteristics (i.e. GC content) or technical (poor quality of the sampled material). If this is the case, we expect to observe a similar impact also on the analysis of SNP-array. We therefore checked for correlation of SNP-array quality scores that refer to noise (MAPD) and fluctuation (waviness) of probes binding (Figure 4.14). However, we could detect a little association of these three scores (Pearson's correlation coefficients between -0.25 and 0.34). We concluded that the CNV calling might be challenged by technical and biological problems that asymmetrically affect SNP array and WES.

Figure 4.14: Quality scores for SNP-array do not correlate with accuracy in our analysis. CNV identified through SNP array technique or WES analyses are not concordant in high quality positions. In the table are reported the Pearson's correlation coefficients for the four comparisons. Following the distribution of patients' values in for the accuracy compared to SNP array quality measures are sparse.

correlation	Accuracy LOSS	Accuracy GAIN
MAPD	-0.25	-0.08
Waviness	0.34	0.23



In conclusion, independently from the tested methods, for some patients it

appeared very difficult to call CNVs. On the basis of the parameters considered, the best method for our analysis appeared to be Control-FREEC for two main reasons: i) though it displays intermediate performances on coverage, it provides good accuracy on our cohort of samples; ii) it is less susceptible to coverage variability. Even if Exome CNV showed similar performances, it always gave an aberrantly high number of variants called. These calls may very likely be false positives because this pipeline misses a normalization step.

4.1.5 The choice of an adequate method to reconstruct clonal composition in tumour samples

At present, the scientific community is lacking a gold standard dataset that would allow assessing the performance of computational tools used for the reconstruction of the subclonal composition of complex cellular populations from genomic data. The ideal dataset would come from a controlled experiment in which the subpopulations present in the sample and their relative abundances are set and known *a priori*, in order to be able to test all the possible sources of error (e.g. capture phase, sequencing phase, bioinformatics pipeline). With the announcement of the ICGC-TCGA-DREAM somatic mutation calling challenge, on Tumour heterogeneity and evolution, a test dataset has been published containing simulated tumour data derived from real tumour data; to date, only a 4 of the 50 tumours are disposable for testing. Since the dataset is not complete, it is difficult to assess methods performances based only on 4 cases, leaving the

need unanswered.

To address this open question, we built a mathematical model for tumour and the corresponding relapse evolution. We, consequently, constructed 90 *in silico* datasets varying standard error and purity of the samples. Standard error was used to simulate the possible fluctuations in the determination of the real VAF of a mutation, because, as already stated above, the sampling of the cell population made through the WES can lead to errors in the VAF estimates. To delineate the impact of VAF variability we introduced standard error at increasing rates of 0, 0.01, 0.05 and 0.09. Of course, also the purity of the tumour samples can affect the clonal reconstruction: we set a purity of 1 to correspond to a tumour sample 100% pure, therefore, without contamination of normal cells. We created 5 datasets for each model with purity ranging from 0.6 to 1.We tested on these datasets four tools for the reconstruction of tumour population, among the most commonly used in the literature. This experimental strategy should allow us to assess which analytical tool is able to better reconstruct the real cellular clonal structure underling the genomic data.

4.1.5.1 Construction of the benchmark *in silico* dataset for clonal analysis

The ideal benchmark should recapitulate:

• the biological conditions: tumour arises in a somatic heterogeneous context and relapse can be the consequence of cells present at exordium and successively evolving;

• the characteristics of the data: AML has generally low mutation rate and, in our

dataset, CNV rates are different in primary and relapse tumours;

• all the possible sources of errors: sequencing errors and purity of the sample can have a great impact on the reconstruction of the clonal population.

Therefore, we decided to proceed building a model with constrains conceptually based on our knowledge.

4.1.5.1.1 Model characteristics

The computational model is defined through a series of matrices and vectors that contain all the parameters that will be used to compute the solution. In particular, we want to describe a situation in which from a normal population of cells, through the accumulation of successive damaging mutations, a tumour expands and, after some time, we are able to capture a timeframe of this evolution, in our case corresponding to the primary tumour sample.

The model is defined by the following parameters:

1. the number of the inspected genomic positions (N): here, we describe in more details only the simplest model that takes into consideration only two genomic positions. However, the model we built consists of a more complicated simulation with 5 inspected genomic positions. In reality, the code is written in order to let the user choose the number of positions to be analysed. It is important to underline that, of course, adding a single genomic position to the simulation increases substantially the computational load for the solution;
2. the code for the identification of the mutation at the inspected genomic position: every position can be in the original state (i.e. not mutated), identified

with a 0, or mutated, identified with a 1;

3. all the possible "genomic" states: the mutational condition of an entire genome is described by a vector of zeros and ones, that reports for every genomic position the presence or absence of the mutation. All the possible states for a model with two genomic positions are:

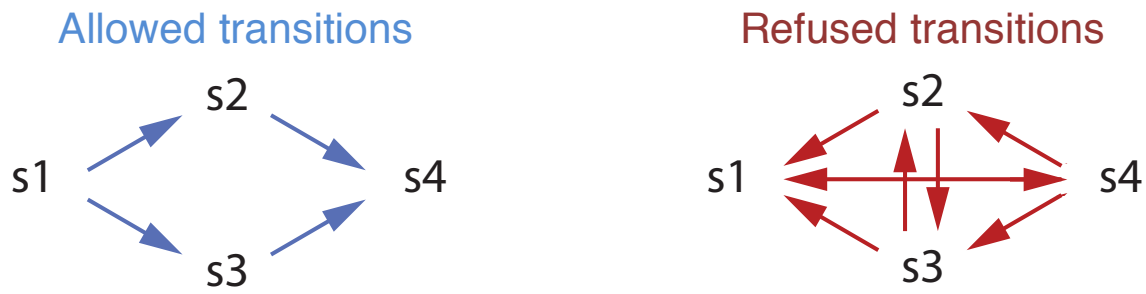
$$Possible\ states = \begin{cases} 0\,0 \longrightarrow s_1 \\ 0\,1 \longrightarrow s_2 \\ 1\,0 \longrightarrow s_3 \\ 1\,1 \longrightarrow s_4 \end{cases}$$

4. the initial population (x_0): considering the primary tumour, the initial population is composed of only normal cells. We arbitrary set the number of cells of the initial population at 100 cells. Because normal cells are all in the s_1 state, the vector is:

$$x_0 = \begin{cases} s_1 = 100\ cells \\ s_2 = 0 \\ s_3 = 0 \\ s_4 = 0 \end{cases}$$

• the allowed transitions: in order to avoid inconsistent situations and to better guide the model to the desired solution, it is necessary to put some restrictions to the transitions among the different mutational states that can be observed. For this reason, we allowed only the transitions that involve the acquisition of a single mutation (the summary of allowed and refused transitions for a two genomic positions model is reported in Figure 4.15).

Figure 4.15: Scheme of transitions allowed in our model. Only transitions indicated with a blue arrow can arise in the model, the occurrence of red transitions is avoided through model constraints.



• the model reactions: on the basis of the allowed transitions, the model is defined with a series of reactions that describes the birth, the death and all the transitions that can take place in the model (see propensity vector for the mathematical definition of each reaction);

• the propensity vector: it defines the actors that will interplay in the model equations. The reactions of changes among states and the associated equations are the following:

$$
\begin{cases}
s_1 \xrightarrow{c_1} 0; \quad c_1 * s_1 = 0 \\
s_2 \xrightarrow{c_2} 0; \quad c_2 * s_2 = 0 \\
s_3 \xrightarrow{c_3} 0; \quad c_3 * s_3 = 0 \\
s_4 \xrightarrow{c_4} 0; \quad c_4 * s_4 = 0 \\
0 \xrightarrow{c_5} s_1; \quad 0 = \dfrac{c_5 * s_1}{(s_1 + s_2 + s_3 + s_4)} * cc \\
0 \xrightarrow{c_6} s_2; \quad 0 = \dfrac{c_6 * s_2}{(s_1 + s_2 + s_3 + s_4)} * cc \\
0 \xrightarrow{c_7} s_3; \quad 0 = \dfrac{c_7 * s_3}{(s_1 + s_2 + s_3 + s_4)} * cc \\
0 \xrightarrow{c_8} s_4; \quad 0 = \dfrac{c_8 * s_4}{(s_1 + s_2 + s_3 + s_4)} * cc \\
s_1 \xrightarrow{c_9} s_2; \quad c_9 * s_1 = s_2 \\
s_1 \xrightarrow{c_{10}} s_3; \quad c_{10} * s_1 = s_3 \\
s_2 \xrightarrow{c_{11}} s_4; \quad c_{11} * s_2 = s_4 \\
s_3 \xrightarrow{c_{12}} s_4; \quad c_{12} * s_3 = s_4
\end{cases}
$$

where c_n defines the constant of association for the reaction (i.e. "death rate",

"birth rate" or "transition rate"), *cc* defines the total number of cells at the previous step, used to impart a spatial constraint to the group of cells that are dividing. The first group of four equations describes the death of the cells in a determined state; the second group of four equations describes the birth of the cells in a determined state; the last group of four equations describes all the possible transitions of allowed mutational states;

- the parameters that define the rates of birth and death: these values are reported in a vector that contains the rates associated to the equations that define the model. The state transition rate is fixed, all the other parameters, instead, depend on the mutations present in that particular state: the state s_0 has death rate and birth rate fixed to 1, every mutation has an associated δ death rate and δ birth rate randomly chosen in these ranges:

$$\delta\ death\ rate\ \in \left[-\frac{1-0.1}{N}, \frac{5-1}{N}\right];\ \delta\ birth\ rate\ \in \left[-\frac{1-0.3}{N}, \frac{4-1}{N}\right].$$

where N is the number of inspected genomic positions in the model.

Note that both these parameters can either be positive or negative and that the presence of two mutations together will not always necessarily result in an increased birth rate or a decreased death rate.

4.1.5.1.2 Definition of the time-points that resemble our set of samples

In order to reconstruct *in silico* the same set of leukemic samples we analysed in practice, which is composed of primary, remission and relapse samples for each tumour, we ran the model with specific parameters and took a snapshot of the

results generated by the model at three different time-points:

- PRIMARY TUMOUR: five genomic positions (N = 5), the initial population is set to 100 normal cells and we let the model run for 30 (t=30) iterations. The primary tumour sample corresponds to the snapshot of the model at time t=30;
- REMISSION: the whole population is formed by 100 normal cells (to resemble the normal state) plus 5 cells that survived the chemotherapy, chosen randomly in the possible states ($N_s(t)$) present in the primary tumour at the end of the simulation (t=30). N_s is the number of states presented at least by one cell in the tumour population at time t. The probability for a state to be represented in each of the 5 cells surviving the treatment is equal to $\frac{1}{\mathrm{Ns(t)}}$.
- RELAPSING TUMOUR: in this case the starting population is the one that has been constructed for the remission; the number of states is five plus the number of not normal states (nns) present at remission (N = 5 + nns). Assuming that it is impossible to loose acquired mutations in a cell, we collapse genetic makeup of cells surviving chemotherapy to a single mutation in the new framework (each corresponding to a nns). In order to avoid transitions between "surviving states" the transition matrix has been modified with additional restrictions. The number of iterations was set empirically to 10 (see next paragraph for the explanation).

To better understand the process described by the model we report in Figure 4.16 a practical example of the three time points. We are analysing a genome that in the primary tumour can present mutations at 5 genomic positions (A, B, C,

D and E). At the end of the first simulation, 3 clones form the primary tumour: s_{29}, s_{30} and s_{31}; the first is antecedent to the other two, having only 3 mutations on the genes C, D and E; the other two evolved differently acquiring respectively a A and a B mutations. $N_s(30)$ for the primary tumour is, therefore, equal to 3 in this case; the 5 cells surviving the therapy will be then extracted with a probability of $\frac{1}{3}$ from the group s_{29}, s_{30} and s_{31}. In fact, we extracted 2 cells for the states s_{29} and s_{30} and 1 cell for the state s_{31}. For evolution of the relapse we, then, launched the simulation on N genomic position where N=5+nns, in our case the number of not normal states at remission was 3 (nns=3), therefore the total genomic positions will be 8. Assuming that it is impossible to acquire 2 independent mutations at the same genomic position, we modified the transition matrix in order to avoid transitions between s_{29}, s_{30} and s_{31}. At the end of the simulation for the relapse sample, we observed that the clone s_{30} expanded overcoming the others and acquired two mutations (G and H) that became dominant plus 3 more subclonal mutations (F, I and J).

Figure 4.16: The output of our model at the three defined time points and reconstruction of the biological framework described. a. For every time-point, we report only the populated states. In the primary tumour we identify three states where s_{29} is antecedent to s_{30} and s_{31} having only mutations C, D and E. At remission all the three states survive at very low frequencies. In the relapse s_{30} expands overcoming the other clones and gaining new mutations. b. A scheme of the biological snapshot at the three time-points with cellular frequencies of the clones carrying the mutations.

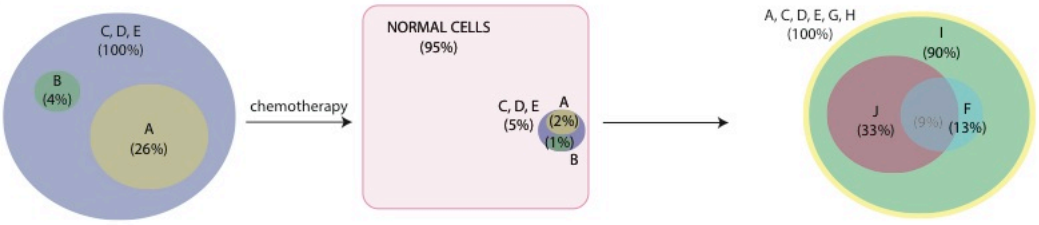
a.

Primary Tumor					
	A	B	C	D	E
s_{29}(70%)	0	0	1	1	1
s_{30}(26%)	1	0	1	1	1
s_{31} (4%)	0	1	1	1	1

Remission					
	A	B	C	D	E
s_{29}(2%)	0	0	1	1	1
s_{30}(2%)	1	0	1	1	1
s_{31}(1%)	0	1	1	1	1

Relapse Tumor								
s_{29}	s_{30}	s_{31}	F	G	H	I	J	
0	1	0	0	1	1	0	0	s_{51} (3%)
0	1	0	1	1	1	1	0	s_{251} (4%)
0	1	0	0	1	1	0	0	s_{179} (7%)
0	1	0	1	1	1	1	1	s_{123} (9%)
0	1	0	0	1	1	1	1	s_{243}(24%)
0	1	0	0	1	1	1	0	s_{115}(53%)

b.



4.1.5.1.3 Setting of the parameters for resembling relapse formation

We did not challenge our model from a mathematical point of view, testing all the parameters in order to choose the better ranges for our framework, because this was not the aim of our study. However, we attempted three combinations of mutation rates and time to relapse to obtain a good output in a reasonable time. Firstly, we set the parameters for the primary tumour and let the model run for 30 steps, considering a mutation rate of 0.05, which is extremely high compared to the real mutation rate of the human genome. These settings were thought in order to increase the probabilities for the model to develop a tumour. We ended up with a population of cells where one state overcame the others, thus resulting in a tumour.

From a theoretical point of view, in order to mimic a physiological situation, it would be better to use a time to relapse smaller than 30 and lower mutation rates. In fact, considering that the primary tumour, in reality, have a mean exordium age around 60, a time to relapse of 30 would be very long compared to the one observed in reality. Indeed, a patient is considered cured if he/she did not develop the disease ten years after the first leukaemia. Therefore, as for the primary tumour, also for the relapse we challenged our model using three different combinations of settings for time to relapse and mutation rates, in order to investigate which conditions better replicates reality, allowing, at the same time, for relapse formation (Table 4.6).

In the first attempt, we used a number of steps for the relapse of 7. We run the model ten times and observed that, with these settings, at the end of the run too many cellular species compose both the tumour and the normal population in the relapse (Figure 4.17.b). Nevertheless, the maximum number of iterations of the model to get a solution showing a tumour population was low (4 iterations). We, therefore, tested the same time to relapse, considering a mutation rate of 0.006. A mutation rate of 0.006 is lower than the one previously tested, but it is still higher than in reality. In fact, it is the rate expected if any mutation occurring in the exome of the cell would behave as a driver mutation, which is clearly very far from reality. In this case the time for simulation increased substantially and the model had to run many times before observing the formation of a relapse tumour (median of 115.5, with a maximum of 1247 iterations, Figure 4.17.b). We needed to run the simulation many times with different set of parameters for birth and

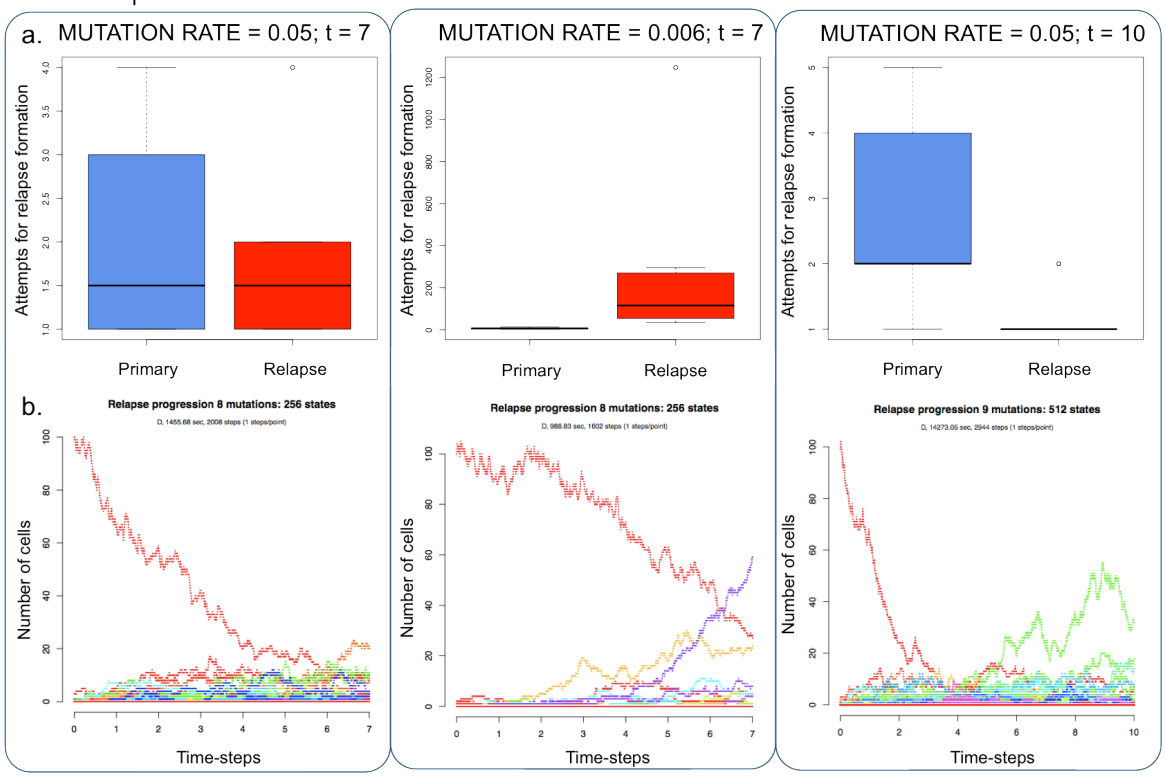
death rates, as explained in paragraph 4.1.5.1.1, because a mutation rate of 0.006 and a t=7 did not allow a cell population to outgrowth the others. Finally, using 10 steps to mimic the time to relapse formation and a high mutation rate of 0.05, we obtained a defined relapse in a reasonable amount of time (Figure 4.17).

Therefore, after these tests, we decided to adopt as final parameters a mutation rate of 0.05 and t=10 and modelled ten couples of primary and relapse tumour using these settings to be used to test several clonal analysis methods tools.

Table 4.6: The parameters tested in our model. The parameters were adjusted in order to find a good compromise between the need to obtain results in a reasonable timeframe and the approximation of the real conditions for development of the disease. Observations report the model output characteristics and the time needed to obtain the results by running the model with the parameters listed.

Relapse steps	Mutation rate	Observations
7	0.05	Relapse is not well defined (many cellular species both in relapse and normal cells)
7	0.006	Mutation rate is still higher than real; very time consuming
10	0.05	Relapse is well defined and can be obtained in reasonable time

Figure 4.17: In our model, different mutation rates and times to relapse (t) largely impact the number of iterations needed to obtain a tumour population and the composition of the observed tumour populations. For every set of parameters, we report: a. the boxplot of the number of times the simulation was ran, resetting the birth and death parameters in order to be able to observe a cellular population in output with tumoural characteristics (the dominant clone harbours at least 3 mutations). We produced 10 primary-relapse couples of samples for every set. Note we report a different scale for the boxplot with mutation rate 0.006 and t equal to 7 to make the graphic more readable; b. an example of evolution of the states during time for relapse. For mutation rate equal to 0.05 and time set to 7 the final snapshot results noisier than for the other two examples.

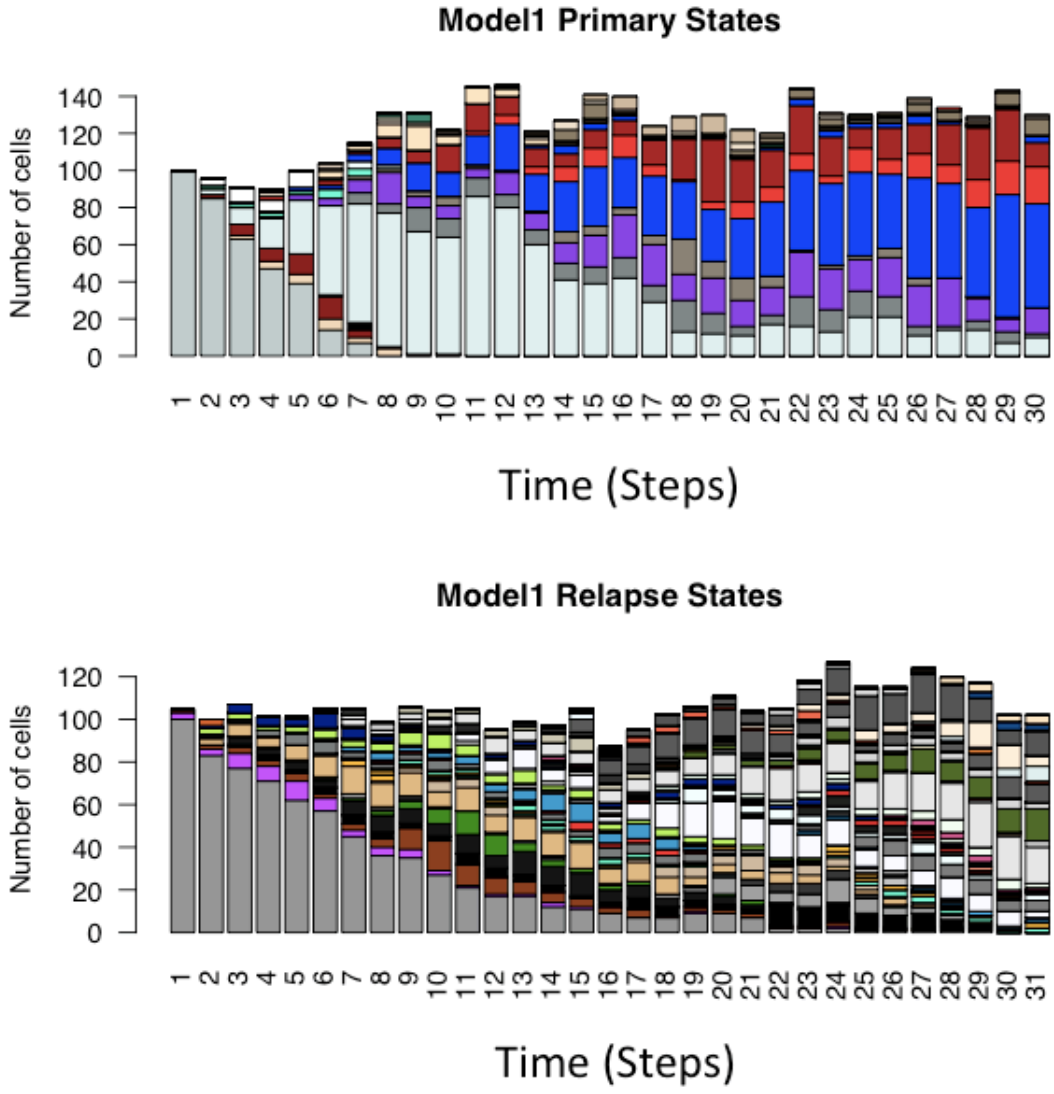


4.1.5.1.4 From model solutions to actual input datasets

The model solutions, described in the previous paragraph (4.1.5.1.3), were calculated in a discrete manner through the Gillespie's method (GillespieSSA R package) and we ran the model resampling the δ death rate and δ birth rate until the expansion of the dominant clone has a minimum number of mutations defined by the user (3 in our case). In Figure 4.18 is reported the composition of the cellular population developed at each time-point during one of the ten

simulations performed with the chosen set of parameters. It is clear that the normal population size decreases gradually in favour of the expansion of the tumour clones. For example, in the primary tumour, the tumour population at the time of diagnosis (t=30) consists of a dominant clone and eight subclones. The relapse, instead, appears more heterogeneous, with more subconal populations compared to the primary tumour, because it starts already with some mutations and can reach a higher number of states.

Figure 4.18: Cellular composition of the primary and relapse populations at each step of our model running: solution 1. Each stacked bar shows the number of cells for all the states (Time) forming the tumour population. Each colour identifies an independent clone and the height of each bar indicates the number of cells for each clone. The first bar (Time 1) represents the normal cellular population and it is composed by the 100% of normal cells, successively the fraction of the normal population decreases and new cells harbouring mutations arise and expand over time.



Once defined the solution of the model, in order to use the information obtained about the tumour populations as a benchmark to test the performances of the different computational methods available for clonal analysis composition, we need to extrapolate the VAFs of the mutations in the primary tumour and in the relapse tumour. As a matter of facts, all the different tools available start, indeed,

from the VAFs and CNVs obtained through WES analysis of the samples, to recapitulate the composition of the tumour population of origin. Therefore, for each of the 10 solutions produced by our model, we produced nine files reporting information on the mutations, their coverage and VAFs, as follows:

• 100 positions of the genome and corresponding base changes are extracted from a file containing all the mutations identified in our cohort of patients (this step was made only to resemble a plausible dataset from the clinical point of view);

• to each position we then associate:

– one of the mutations in the model (M), randomly;

– the number of DNA copies present in the primary tumour exactly at that position (CN_{PT}). This number is extracted with a *per* base probability distribution determined by the analysis of all the CNVs detected in our samples; if this value differs from 2, then there is a random determination of the number of copies that are mutated (MCN_{PT})

– the number of DNA copies present in the relapse tumour (CN_{RT} ; as before, the probability distribution in the relapse is slightly different in this case)

– the error rate at that position for every "coverage" (err)

– the coverage in the three samples (per primary tumour $cov_{PT,}$ per remission cov_{R}, per relapse tumour cov_{RT}), extracted in the range [60-150X] as the majority of the positions in the samples we analysed present the same distribution (we avoided extreme coverage because sometimes it is associated to a bias in the capture or in the alignment phase)

• once that CNs, error and coverage were set for every position, we computed the number of reads for the following categories: reference (R) and alternative (A) coverage for the primary tumour, for the remission and for the relapse, respectively with the equations:

$$PT_R = \left|\left[1 - \left(\frac{VAF_{PT}(M) * MCN_{PT} * purity}{CN_{PT}} + err_1\right) * cov_{PT})\right]\right|$$
$$PT_A = \left|\left[\left(\frac{VAF_{PT}(M) * MCN_{PT} * purity}{CN_{PT}} + err_2\right) * cov_{PT})\right]\right|$$
$$R_R = \left|\left[\left(1 + err_3\right) * cov_R)\right]\right|$$
$$R_A = \left|\left[\left(0 + err_4\right) * cov_R)\right]\right|$$
$$RT_R = \left|\left[1 - \left(\frac{VAF_{RT}(M) * MCN_{RT} * purity}{CN_{RT}} + err_5\right) * cov_{RT})\right]\right|$$
$$RT_A = \left|\left[\left(\frac{VAF_{RT}(M) * MCN_{RT} * purity}{CN_{RT}} + err_6\right) * cov_{RT})\right]\right|$$

Purity (purity) and error (err) are alternatively fixed to a neutral value, in order to create 9 datasets for every solution: 4 will have purity fixed to 100% and increasing error rate equal to 0, 0.01, 0.05, 0.09 or, vice versa, 5 will have error rate fixed to 0 and values of purity growing from 60% to 70%, 80%, 90% and 100%.

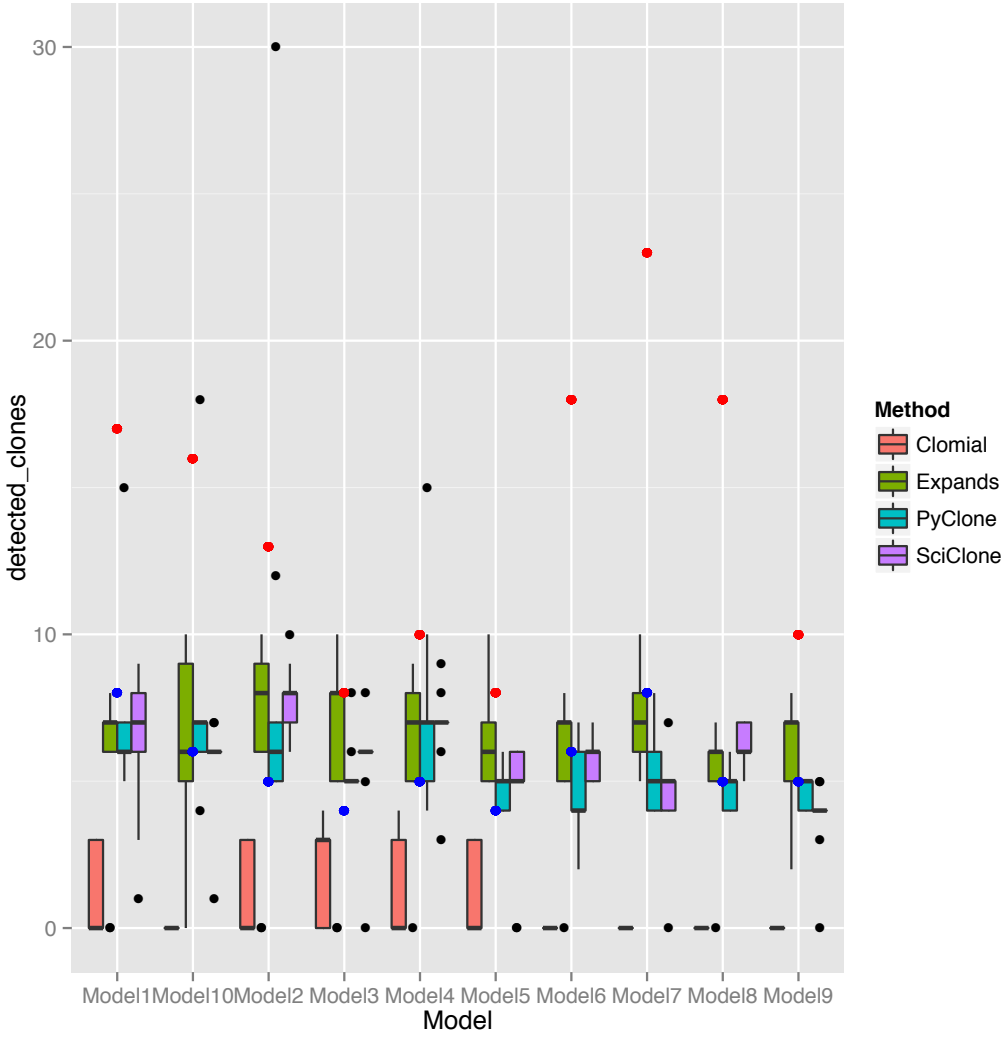
4.1.5.2 PyClone is the best performing clonal analysis decomposition tool on our benchmark

We have now created a benchmark composed of 90 files, reporting observations referred to 10 tumour models based on varying error rate and purity of the sample. At this stage, we analysed all 90 datasets with four of the most used tools

for clonal reconstruction: Clomial, ExPaNds, PyClone and SciClone. As stated above, all these tools start from VAFs of single point mutations and CNVs data, in order to reconstruct the clonal composition of a tumour, generating as an output the number of clones forming the tumour population, their associated frequencies and the variants belonging to each clone or subclone. We could, therefore, evaluate the performance of each tool, comparing the results of the analysis of our benchmark dataset obtained with each tool to the model's solutions.

The first and easier task to be evaluated is the identification of number of clones. In Figure 4.19, for every tool, we show the total number of clones forming the tumour population as a red dot and the number of detectable clones (i.e. clones with VAF higher than 5%, i.e. VAF of heterozygous mutations 2.5%) with a blue dot. For every method, the boxplots are associated to the number of clones identified varying error rate and purity. Clomial always underestimates the number of clusters, failing in the assignment of clone identification. In contrast, Expands, PyClone and SciClone tend to approach the real number of clones.

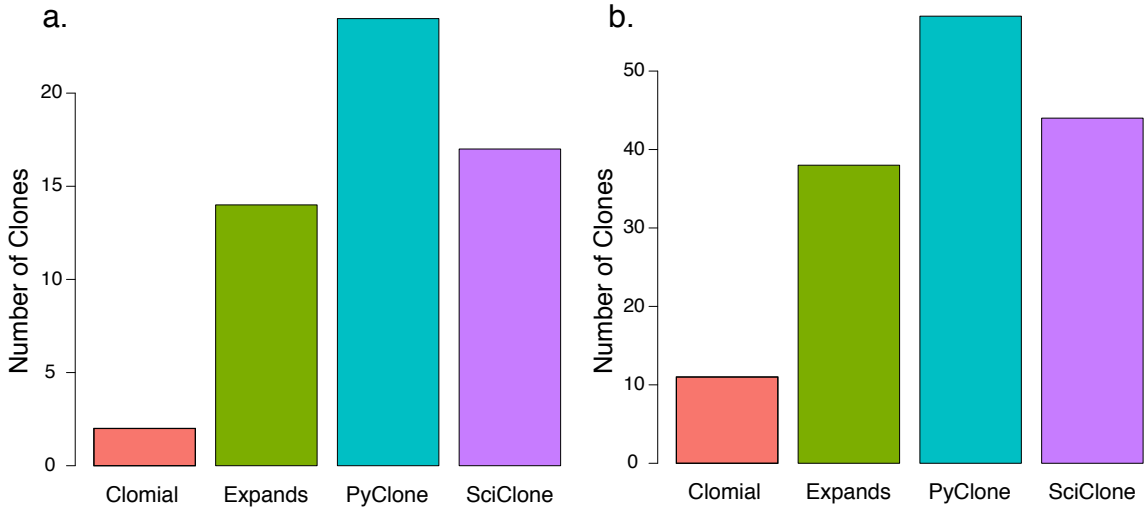
Figure 4.19: Performances of four clonal composition analysis methods on our benchmark dataset. For each model and method is reported a boxplot of the number of clones identified varying error rate and sample purity. The total number of clones is represented with a red dot; the number of detectable clones (frequency > 5%) is depicted with a blue dot. The outliers are depicted as black dots. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



We, next, considered as reference for the number of clones, only the exact number of detectable clones in the model (i.e. clones with frequency >5%). We decided to exclude low frequency clones from further evaluations because, at low frequencies, it becomes really difficult to discriminate groups of mutations with similar behaviour. Moreover, we mostly care about grouping higher frequency mutations, because the assignment of low frequency mutations to subclones, though they might play a determinant role in relapse expansion, is a challenging

problem starting from WES data. Indeed, even mutations discovery at low frequency is challenging from WES data and the ones we detect are selected by chance from the low frequency mutations population. We used the reference number of clones to evaluate the number of times each method identifies the exact number of clones (Figure 4.20.a) or misses the exact result by maximum one clone (Fig 4.20.b): the higher is the bar associated to the method, the better it performs in the identification of the exact number of clones. PyClone results the best performing method, grasping the correct numerosity in 24/90 cases (27%) and missing the correct results by maximum one clone in 57/90 cases (63%). SciClone ranks after PyClone with 17 (19%) and 44 (49%) exact and almost exact results, respectively; Expands ranks third (14 and 37) and Clomial gives the worst results, catching at uttermost the 12% of the correct solutions (2 exact results and 11 missed).

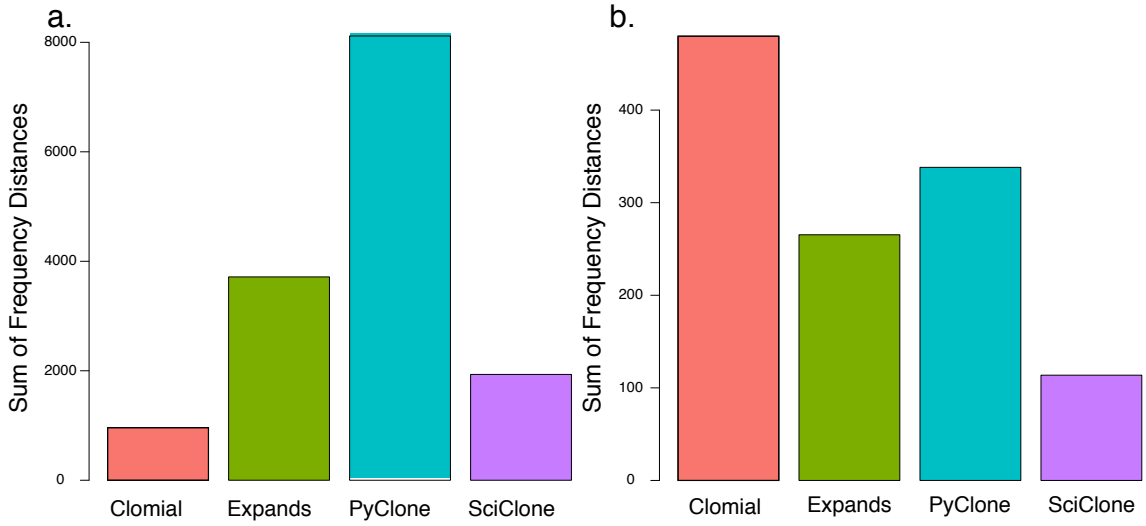
Figure 4.20: Evaluation of the performance of the different methods in discerning the right number of clones. On the y-axis is reported the number of times each method is able to determine the exact number of clones in the tumour population (a), or misses the correct result by maximum one clone (b).



Then, in the cases in which the number of identified clusters was right, we

evaluated whether also the frequencies associated to each cluster approximate the real frequency. We calculated the distance of the frequency of each real clone from the nearest frequency identified by the tested tools. When the methods identified one clone less or one clone more than the real ones, we simply added its distance to the sum. Figure 4.21 shows, for each method, the sum of distances both not normalized (Figure 4.21.a) and normalized for the number of exact calls (Figure 4.21.b). Considering this parameter, SciClone is the tool that performs the best.

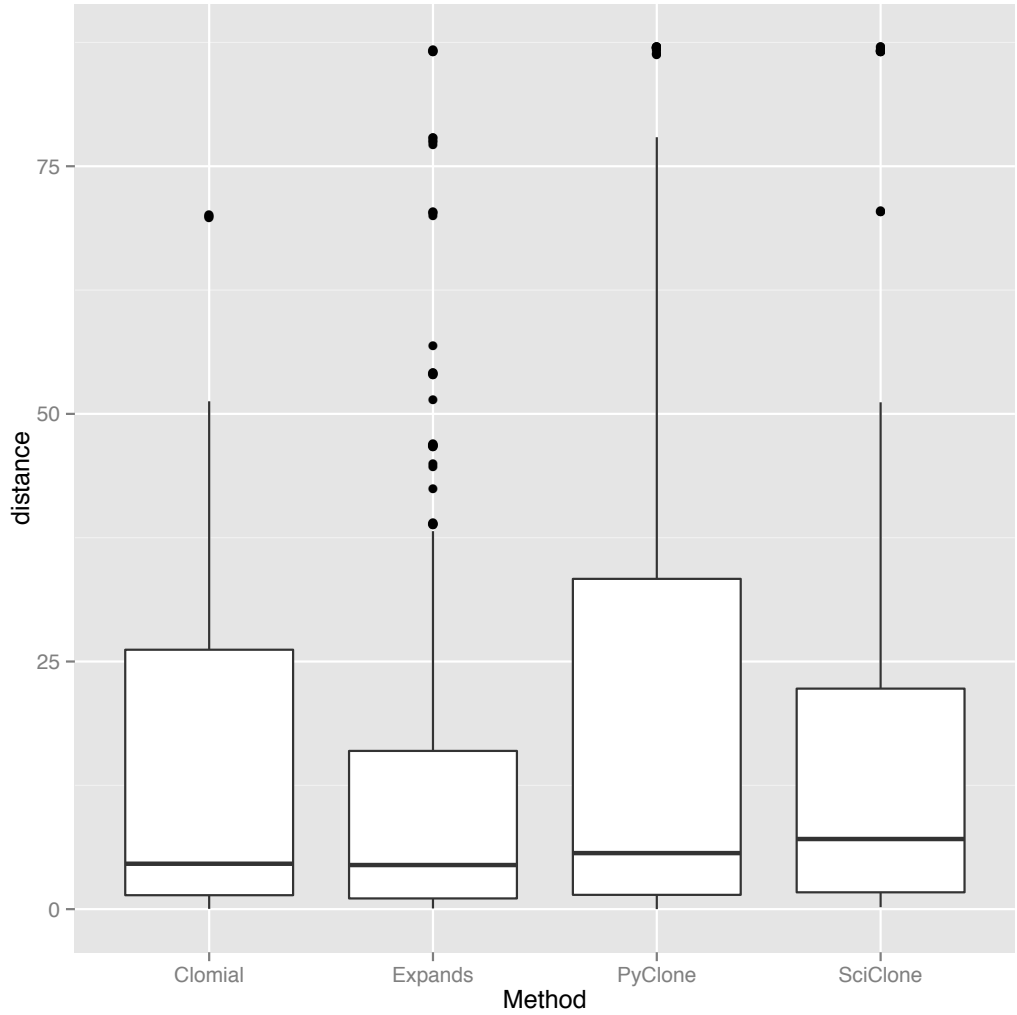
Figure 4.21: Evaluation of the performance of each tool in the determination of clonal frequencies. For each method, when the number of clones identified was correct (plus or minus one), we computed the sum of the distance of the frequency of each real clone from the nearest frequency identified by the tested tools. Panel a. shows the absolute distance, panel b. the distance normalized for the number of exact calls. The impact of the normalization step becomes clear for Clomial, which fails to retrieve the exact number of clones in the majority of the cases and appears to have little absolute distance. However, after normalization, becomes the worst performing in the prediction of frequency.



At first sight, the immediate conclusion from these results would be that SciClone gets closer to the real frequency, however, this is not correct. In this type of measurements mistakes are not uniform. Indeed, showing directly the distributions of the distances of the predicted frequencies from the real ones for

each method, it is clear that the method, in general, getting closer to the real frequencies is ExPands (Figure 4.22). In fact, in most cases, ExPands makes little mistakes and rarely misses completely the result. SciClone instead, has the bigger median error but it is less prone to big mistakes. PyClone, finally, has a median distance lower than SciClone but it allows for the biggest distances from the correct frequency.

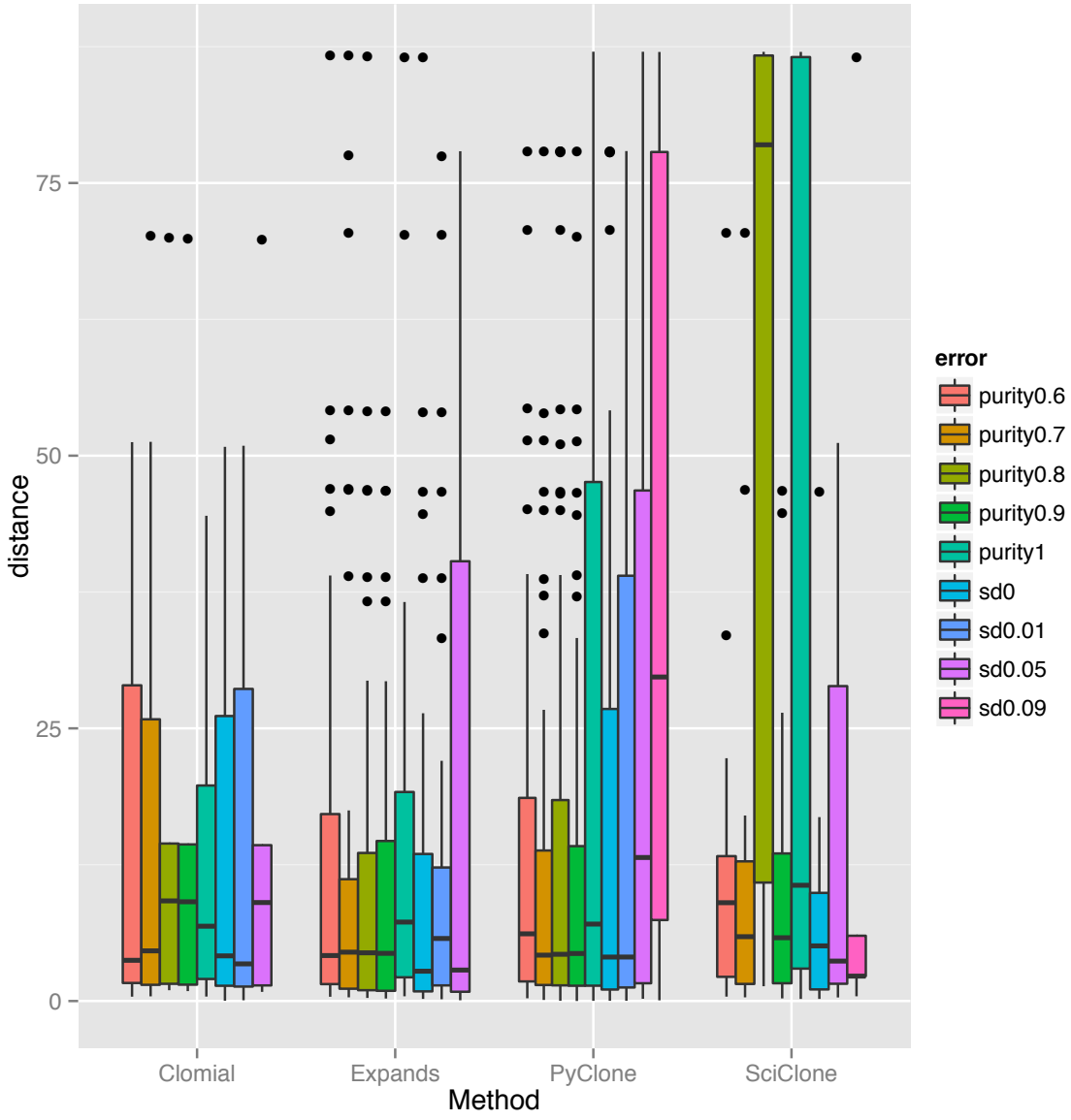
Figure 4.22: Distances from the correct number of clones grouped by method used for clone identification. The distribution of the distances of the frequency from the real reveals that SciClone, the method having overall little distances, counter intuitively has the highest distance values in the majority of cases. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



In 70 of the 90 datasets produced for testing of the methods, we introduced

errors in the VAFs proportions, emulating an altered purity of the sample of origin or a deviation from the effective VAF of the variants. We, therefore, decided to quantify the effect of these errors on the performances of the tools, measuring again the distances from real frequencies. Indeed, the responses of the different methods to data variability are diverse (Figure 4.23). Datasets without noise are not always the easiest to deconvolute: SciClone, for example, has the broadest boxplot for the case with purity 1 (and standard deviation 0). We noticed that the specific dataset used for the analysis has a great impact on the performances of the methods, because two separate samplings with equal characteristics (i.e. purity 1 and standard error 0) always give different results. PyClone is highly affected by changes in the standard deviation and we observed growing distances associated to bigger standard deviations. For SciClone, Expands and Clomial is even impossible to produce the boxplots for some groups, indicating that they never get close to the exact number of clones for that specific condition. The best performing tools are Expands and PyClone, the former is more robust to all sources of error, the latter slightly outperforms the others in ideal conditions: sd < 0.05 and purity > 60%.

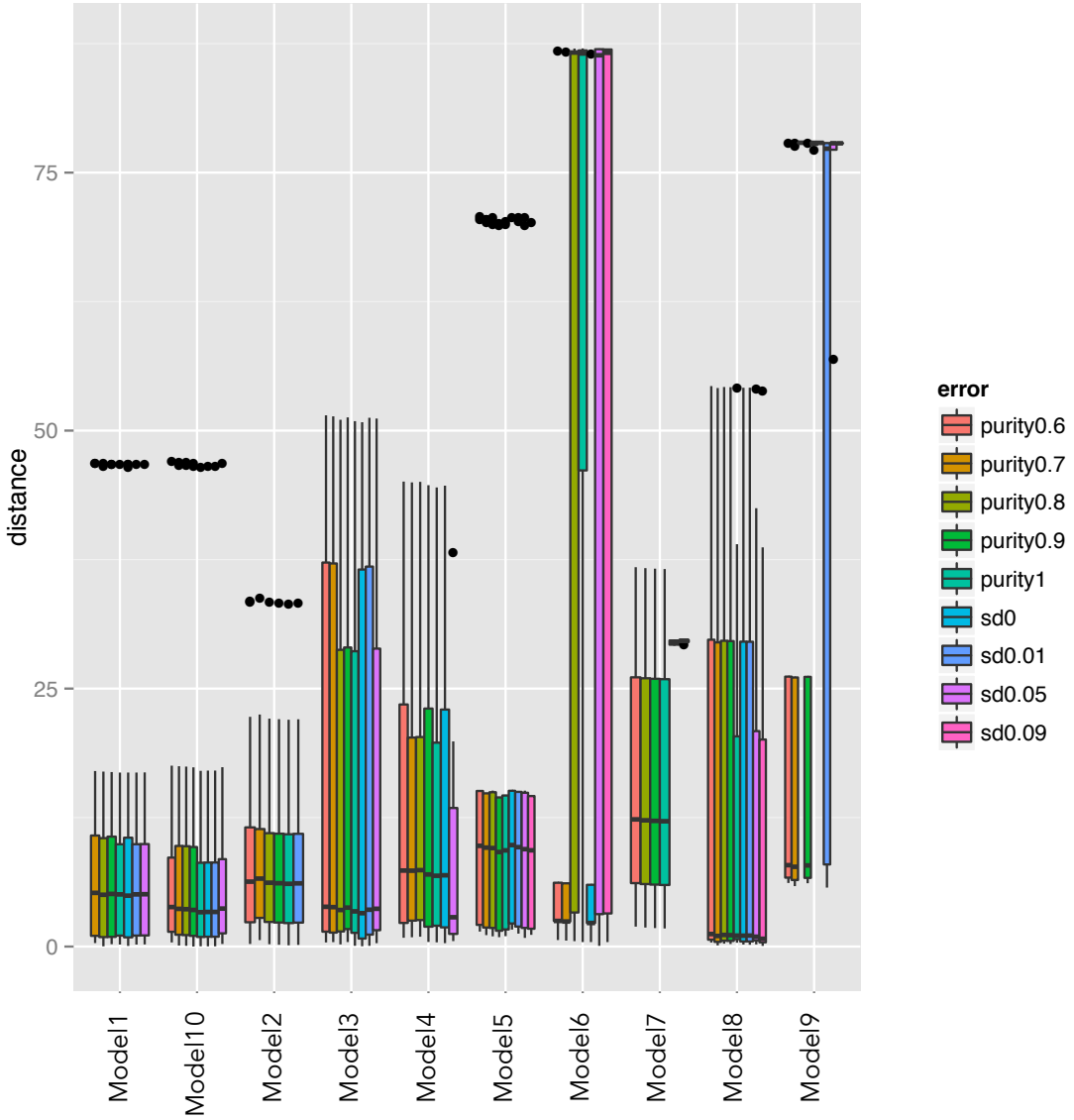
Figure 4.23: The impact of external sources of variation on the capacity to discern clonal composition. For every method the boxplot shows the distances from the real frequencies computed under different sources of noise. Expands is the more robust to external variability; on the contrary, SciClone has a fluctuating behaviour and PyClone is influenced by the standard deviation. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



Despite the source of variability in the dataset, a major challenge in clonal decomposition is played by the inner characteristics of the clonal population. Grouping together the results analysed with the four methods for each model derived in the construction of the benchmark, it is evident that the error has a minor effect compared to the variability among solutions (Figure 4.24). Indeed,

some models are easier to be decomposed, for example model 1 and 10 have smaller boxplots, indicating that they have little median distance of the frequencies from real. On the other hand, model 6 and 8 have broader boxplots for the distances, probably reflecting a complex clonal composition.

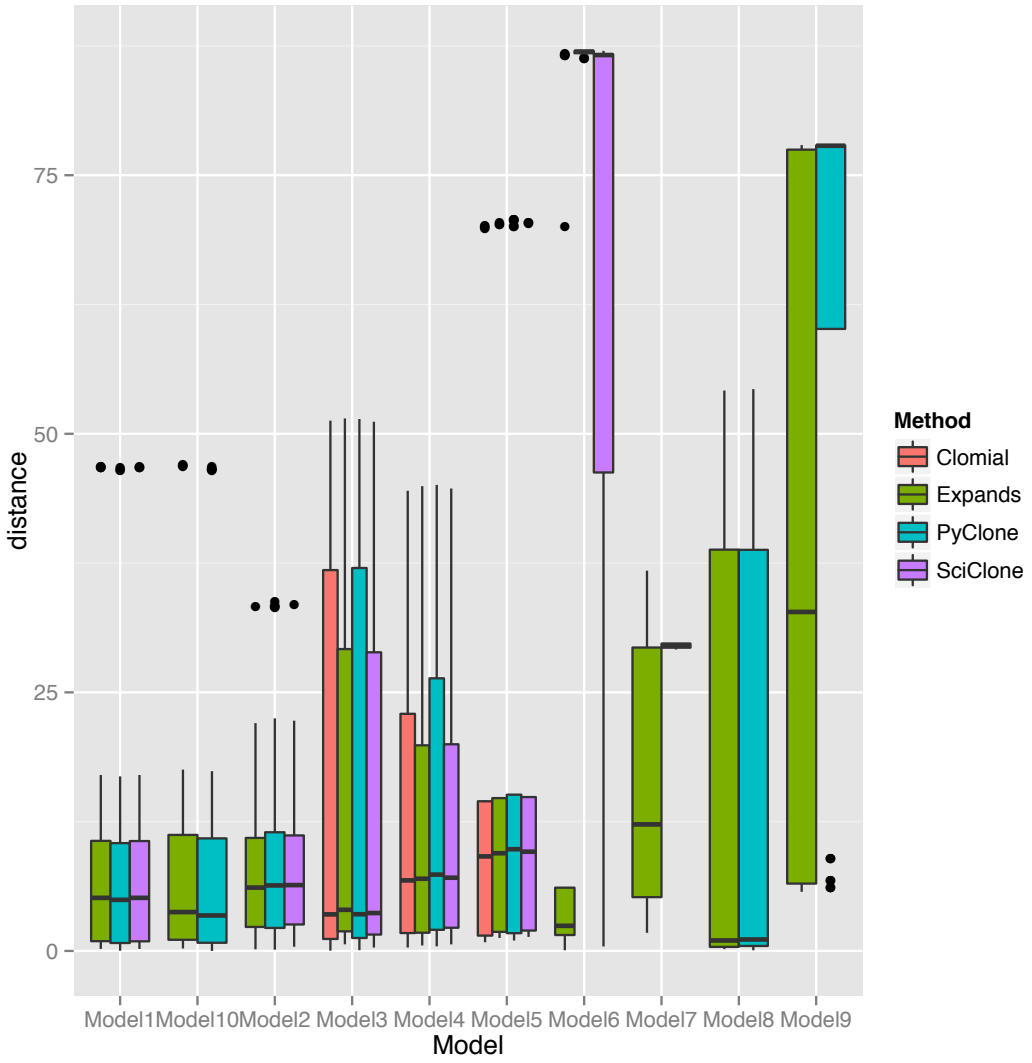
Figure 4.24: Boxplots of the distance from exact frequencies for couples solution-error source. Regardless of the method used for the determination of clonal composition, we studied the association between errors and models and observed that the model have a great impact in the solution of the problem. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



In particular, analysing the distances in function of the models and the methods used for calling the tumour subpopulations, it is clear that all the methods behave

similarly in simple cases (models with a low number of clones or populations with very different frequencies; Figure 4.25, Model1 and Model 3). However, their behaviour is very different for the difficult cases (models with a complex clonal composition and where different clones have similar frequencies in the tumour population) and, in many cases, the methods fail even in identifying the number of clones present in the population (Figure 4.25, Model6 and Model9). Expands looks more robust, though we know that for standard deviation equal to 0.9, it is unable to produce results. PyClone always returns a result but, in difficult cases, it is far from the real value.

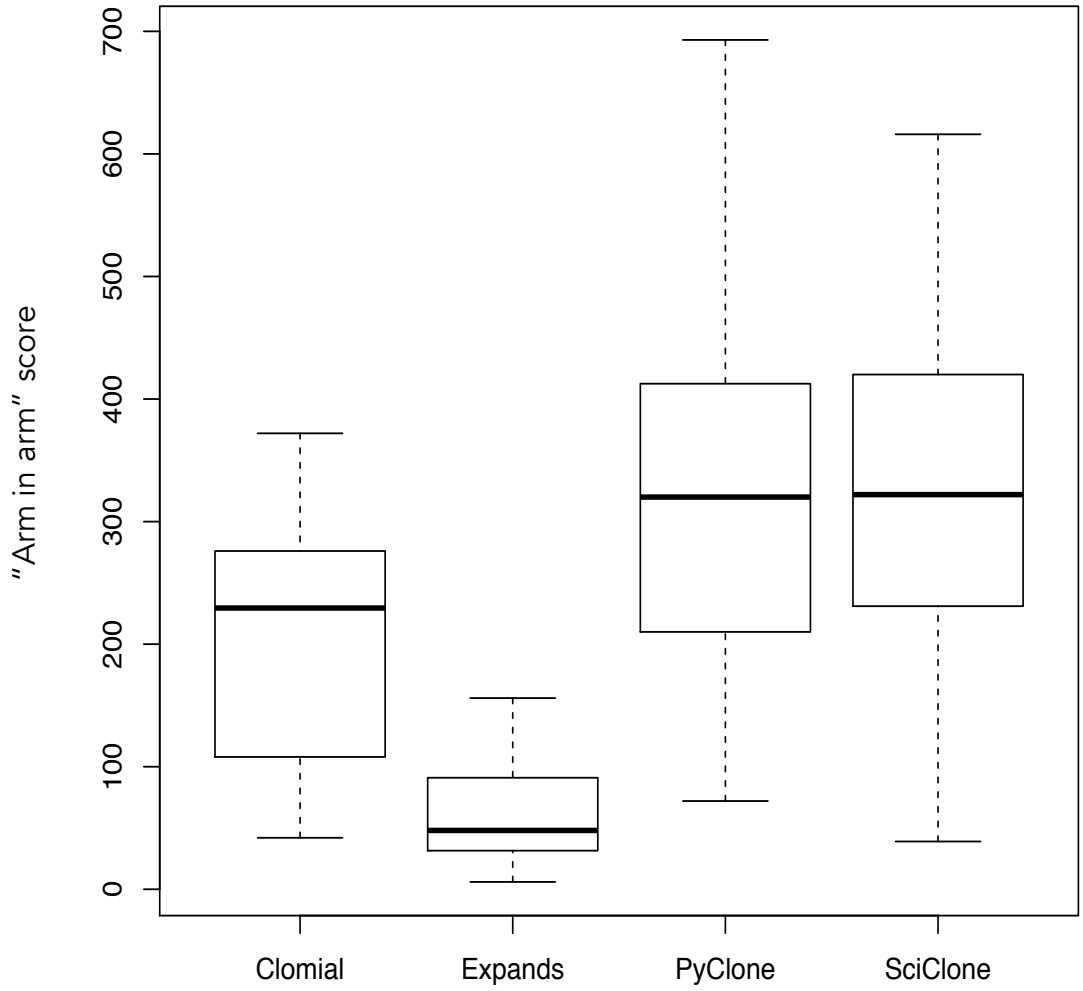
Figure 4.25: Robustness of the methods considering different complexity of the models. A boxplot of distances is reported for every couple model tool. The boxplot is produced only when the tool was able to ascertain the correct number (plus or minus one) of clones forming the population. For simple models the responses of the algorithms are very similar; complex situations show very heterogeneous results. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



The last property to assess with our analysis was the ability of the different tools to correctly group mutations, or, in other words, the ability of a tool to assign to the same clone the variants belonging to the same cell in the tumour population. In order to evaluate this ability, we used an "arm in arm" score build, multiplying the number of mutations in the N most populated intersections (where N is the correct number of clones for that model) for the number of clones identified. The

higher the "arm in arm" score is, the better the algorithms perform, because we consider the most populated intersections as correctly identified clones (i.e. regardless of the frequency assigned by the methods to each clone, we evaluate whether the methods are able to group in the same clone mutations that originally were together). Since it is more difficult to get a high intersection when the number of clones is high, we multiplied the value for the real number of clones. Surprisingly, Expands gives the worst results. Even if the number of clones and the frequency of clones identified by Expands are correct, the mutations are wrongly grouped (Figure 4.26). PyClone is slightly better than SciClone, but, concerning this aspect, the performances of the two methods are comparable.

Figure 4.26: "Arm in arm" score for the four methods. The boxplots reports the "arm in arm" score distribution for all the cases in which the method correctly identifies the number of clones. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



To summarize, a good method for clonal decomposition in a tumour population should be able to: i) determine the right number of clones, ii) determine the right frequency of the clones and iii) correctly group together the mutations belonging to each clone. From our analysis, performed using our benchmark dataset, PyClone emerged as the best compromise for our need.

4.1.5.3 Testing clonal analysis decomposition tools on a public dataset we obtain poor results

The actuality and the difficulty of the deconvolution of tumour population is

underlined by the fact that the DREAM challenge organizers decided to dedicate a specific challenge to the Tumour heterogeneity and Evolution. The challenge results have not been published yet, however, the organizers made publically available a small training dataset, of which we took advantage to test the different methods. The dataset is composed by two tumours for which they provide: the files with the list of mutations (SNVs) and CNVs, information on the real purity of the sample, the number and frequency of the real clones and the list of mutations belonging to each clone. On this specific dataset we were able to test only Expands and SciClone because PyClone needs more than one sample for the same patient to work and Clomial does not give an output. Nor Expands neither SciClone predicts the correct number of clones. The distances from the correct number of clones are 3 and 6 for SciClone and 5 and 6 for Expands (Table 4.7).

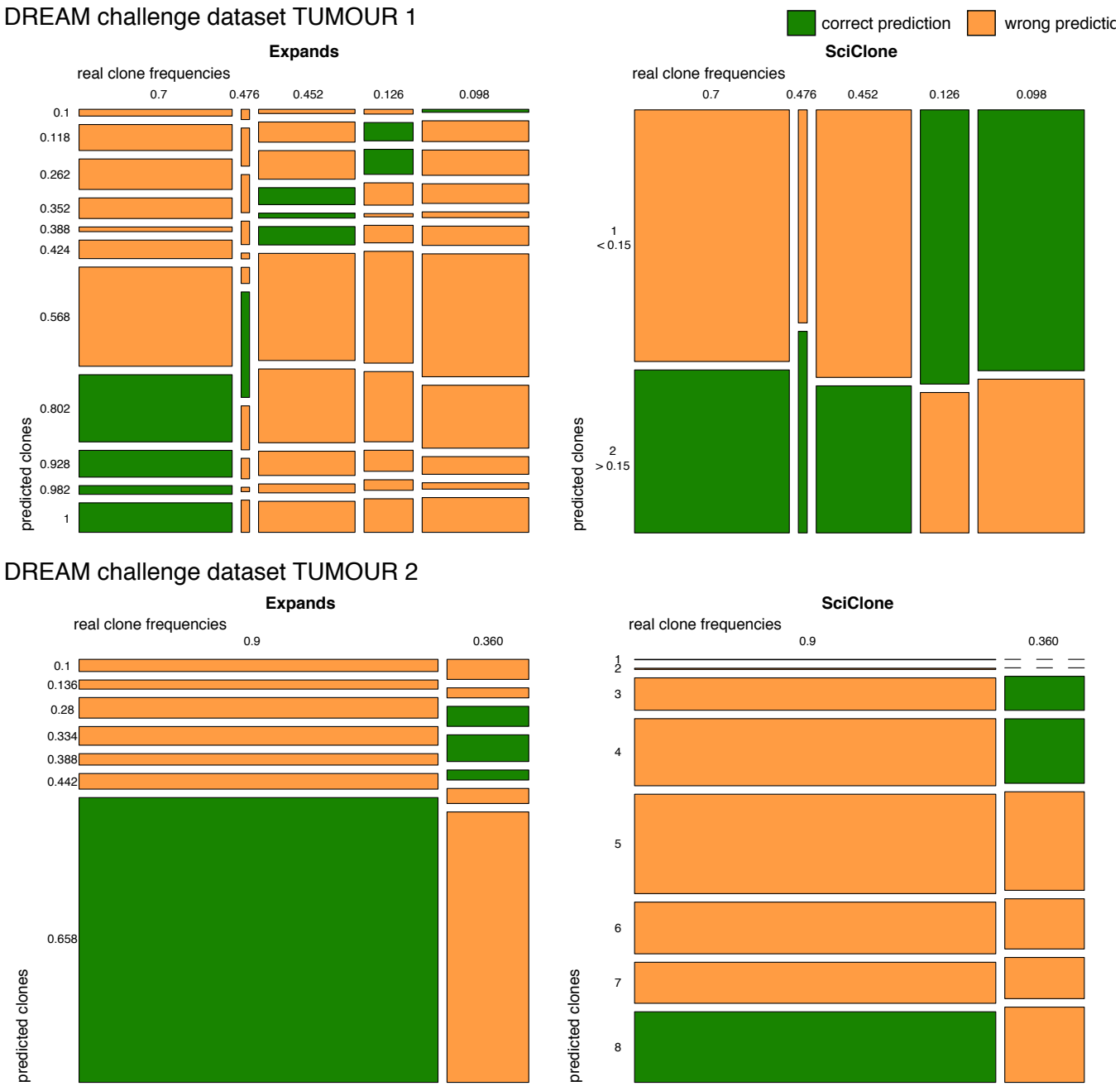
Table 4.7: Number of clones identified using the DREAM challenge datasets.
Number of clones identified by Expands or SciClone, compared to the numbers provided by the DREAM challenge (Real).

Dataset	Real	Expands	SciClone
Tumour 1	5	11	2
Tumour 2	2	7	8

For every mutation we disposed of the cluster of membership and its frequency within the cluster, therefore, we were able to plot the correspondence between predicted and real clusters. We plotted the numbers of mutations as squares, where height and width of the squares are proportional to the number of mutations with those characteristics in the predicted and the real clones, respectively. Expands gives in output the frequency associated to the predicted

clones, while, for SciClone, we were able to identify a discrimination criterion. Therefore, we were able to determine which mutations were assigned to the right clusters (green squares in Figure 4.27). In many cases the real clusters were split or grouped in the prediction: for example, on the dataset tumour 1, Expands splits a single real clone at 0.7 frequency in 4 clones with frequency ranging from 0.8 to 1 (we consider these results correct because the purity of the sample in tumour 1 was 0.7); on the other hand, SciClone groups together at frequency higher than 0.15 three clones with real frequencies ranging from 0.452 to 0.7. Even worst, excluding the correct predictions, we noticed that mutation belonging to one clone could be distributed among all the clones in the predictions (yellow squares in Figure 4.27). Since the proportion of yellow squares (i.e. wrong predictions) in the graph is considerably larger than the proportion of green squares (i.e. correct predictions), we conclude that the performances of SciClone and Expands on this dataset are not satisfactory.

Figure 4.27: Clone prediction on the DREAM challenge datasets. For each tumour dataset (TUMOUR 1 and TUMOUR 2) and for each method (Expands and SciClone), we compared the predictions for every mutation with the real clone. The square sides are proportional to the number of mutations belonging to each specific group. Green squares are mutations correctly classified; yellow squares are associated to wrong classifications.



Reporting the results obtained on the same datasets by the teams that participated to the DREAM challenge we observed that the tested methods, compared to the group of publicly available ones we reported, perform better in the identification of the number and proportion of subclones in the tumour

population (Table 4.8). On the other hand, the performances in the task of mutation assignment to subclones remain poor with the only exception of the team GuanLab_SMCHet on TUMOUR2.

Table 4.8: Performances of the teams that participated to the DREAM challenge on Tumour1 and Tumour2. We reported the preliminary results obtained by the teams in their pre-submission round on the dataset previously described. For the three sub-challenges reported (predicting number of clones, predicting subclone proportions and determining mutation assignment to subclones) a score of 1 is associated to perfect performances and measures respectively the difference between true and inferred number of clones, the mean absolute difference obtained comparing true and predicted cellular prevalence and the correlation between the true and predicted matrices clustering together the mutations belonging to the same clone.

'UMOUR1	**Predicting Number of Subclones**	**Predicting Subclone Proportions**	**Determining Mutation Assignments to Subclones**
GuanLab_SMCHet	0.67	0.96	0.47
'eam Markowetz	0.67	0.91	0.19
'he overtaker	0.83	0.85	0.07
MC-Testers	0.33	0.69	0.05
acktumour	0.5	0.64	0.05
GISL	NA	NA	NA
'UMOUR2	**Predicting Number of Subclones**	**Predicting Subclone Proportions**	**Determining Mutation Assignments to Subclones**
GuanLab_SMCHet	1	0.98	0.87
'eam Markowetz	0.67	0.93	0.61
'he overtaker	0	NA	-0.49
MC-Testers	0.67	0.39	0.26
acktumour	0.67	0.39	0.26
GISL	1	0.62	-0.06

4.2 Biological results

The refinement of the methods described in the first part of our thesis had the main scope to have at our disposal the cutting edge bioinformatics tools that best adapted to the specific needs of our study. Our aim was, indeed, to analyse in depth, from the exomic point of view, a cohort of 30 patients affected by AML, in three phases of evolution of the disease: primary tumour, remission and relapse. We concentrated our efforts, in particular, on delineating the commonalities and the differences among our groups of samples, concerning single nucleotide, copy number variants and small insertion/deletions (indels).

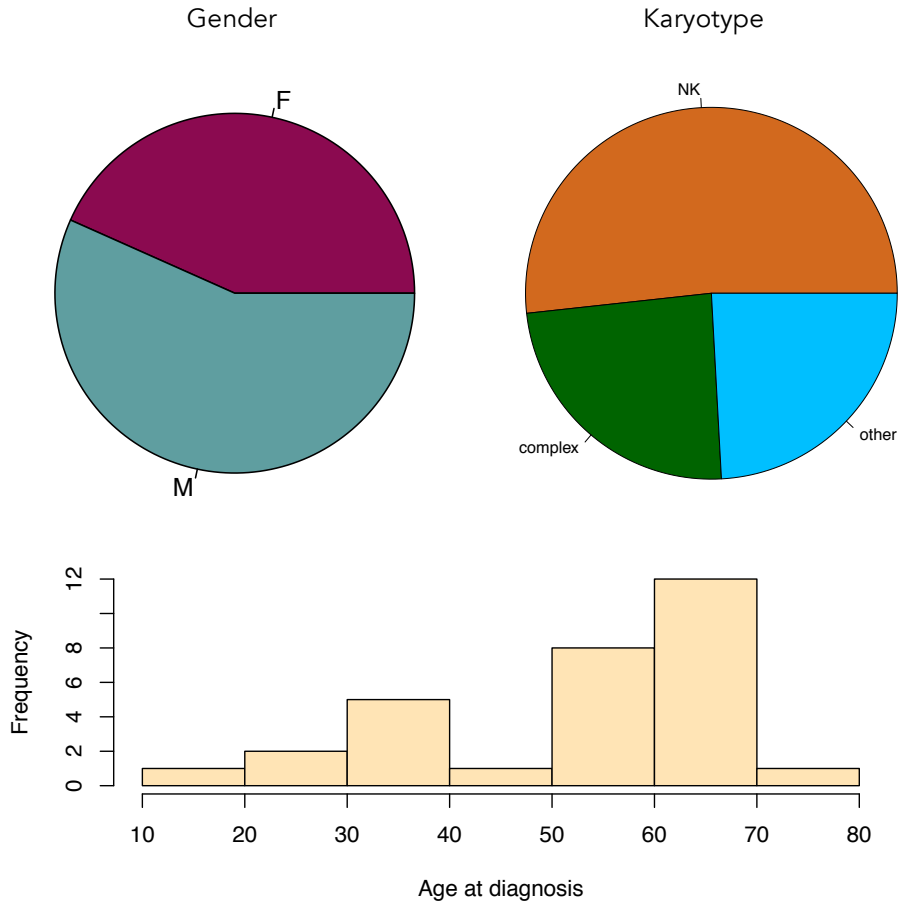
4.2.1 Patient's characteristics

The general characteristics of the patients of our cohort are reported in Table 4.9 and schematized in Figure 4.28. Our cohort was balanced for gender: males and females were respectively 17 (57%) and 13 (43%). Half (15/30) of the patients presented a normal karyotype; all the others presented either complex karyotype (7/30, 23%) or specific chromosomal rearrangements (excluding one patient for which the karyotype information was not available). The majority of patients were diagnosed at ages between 50 and 70 years old (67%), however the range was between 18 to 73 years of age.

Table 4.9: General characteristics of the patients collected for our study. For every patient in our cohort we report gender, age at diagnosis and karyotype.

PATIENT	GENDER	AGE	KARYOTYPE
BO1	M	32	NK
BO2	M	42	(+8);t(2;10)(q33;p13)
BO3	F	34	NK
BO5	F	55	complex
BO6	F	62	complex
BO7	M	57	NK
BO8	F	70	complex
BO9	M	53	complex
BO17	F	64	complex
TO1	F	67	NK
TO2	F	73	NK
TO3	M	58	NK
TO4	M	67	46, XY, +11
TO6	M	70	t(8;21)
TO7	M	70	NK
TO8	F	61	NK
UD1	M	33	(-6,-11)
UD2	M	39	complex
UD3	M	55	NK
UD4	F	59	complex
UD5	M	67	NK
UD6	F	18	t(8;21)
UD8	M	28	NK
UD9	F	32	NK
UD10	F	22	INV16
UD11	M	62	NK
UD12	M	57	t(10;11)
UD13	F	54	NA
UD14	M	67	NK
UD15	M	68	NK

Figure 4.28: Characteristics of our patient's cohort. Distribution of our patients for gender (left pie chart), karyotype (right pie chart) and age at diagnosis (histogram).



Concerning the clinical information, the risk stratification and the follow up of our patients, the data are reported in Table 4.10. The diagnosis was performed cytogenetically in 4 cases, molecularly in 11 and with both techniques in 4 patients; only in one case it was done ematologically and for 10 cases we did not have this information. The majority of our AMLs belong to the FAB classification M1 (7/30, 23%); 3 AMLs were M0-M1, 5 were M5 and 4 M4; the remaining were M0 (3), M2 (2) and secondary (WHO classification, 2) (for 4 patients the FAB subtype was not available). We received the information about risk stratification for 25 patients: 13/25 (52%) had a high risk of relapsing, 3 (12%) intermediate risk and 9 (36%) standard risk. The treatments were heterogeneous, the doses and number of cycles varied a lot among patients, but the drugs used, in the vast

majority of the cases, were Fludarabine, Cytarabine, Idarubicin and Etoposide. Moreover, the post-induction strategies were adapted to patient's response and characteristics. Finally, concerning the follow-up, we know that only 10 (33%) patients achieved second complete remission, while, unfortunately, the rest died of this pathology or complication related to transplants.

Table 4.10: Clinical information of our cohort of patients. For every patient, we report the FAB subtype detected at the first exordium of the tumour, the risk category (S: standard risk, I: intermediate risk, H: high risk), the type of diagnosis, the induction therapy, the consolidation therapy and, finally, the follow-up (CR, complete remission; D, deceased). The abbreviations used for treatments are the following: FLAI=Fludarabine+Cytarabine+attenuated-doses of Idarubicin; IDA=Idarubicin, ARAC=Cytarabine; 3+7=standard treatment; AZA=Azacytidine; E=Etoposide; DAUNO=Daunorubicin; HD=high doses; LD=low doses. NA, not available.

Patient	FAB	RISK	Diagnosis	Induction therapy	Post induction 1st line	Post induction 2nd line	Post induction 3rd line	Follow-up
BO1	M1	S	NA	FLAI + IDA ARAC (X2)	auto-transplant	allo-transplant (TMO)		CR
BO2	M5	H	NA	FLAI + IDA ARAC	HD ARAC	velcade+zarnestra		D
BO3	M1	H	NA	FLAI	HD ARAC	auto-transplant	allo-transplant	D
BO5	M4	H	NA	FLAI+E	HDAC IDA	BUS/ALKERAN	auto-transplant	D
BO6	M1	S	Cytogenetic	3+7	1--7	auto-transplant		CR
BO7	M0-M1	I	Cytogenetic	3+7 (X2)	1--7	HD ARAC	allo-transplant	CR
BO8	M0-M1	H	Cytogenetic	NA	1--7			D
BO9	NA	NA	Molecular	DAUNO+E+FLAI	auto-transplant			D
BO17	M0	H	Molecular	AZA				D
TO1	M5	H	Ematologic	IDA +ARAC +E	IDA+ARAC			D
TO2	M5	H	NA	IDA+ARAC	IDA+ARAC			D
TO3	M1	NA	NA	IDA +ARAC +E	HD ARAC			D
TO4	NA	NA	NA	IDA +ARAC +E	allo-transplant (sibling)	allo-transplant (sibling)	OCN AZA	CR
TO6	NA	NA	Cytogenetic	IDA +ARAC +E				D
TO7	M4	H	NA	IDA +ARAC +E	Depakin, Rocaltrol, Aisoskin			D
TO8	NA	S	NA	IDA +ARAC +E				D
UD1	M0	H	Molecular	FLAI + IDA ARAC	HD ARAC			CR
UD2	M2	S	Cytogenetic and Molecular	FLAI	ARAC+IDA(1); HD ARAC(1) ARAC + DAUNO (1); HD	allo-transplant		CR
UD3	M5	S	Molecular	ARAC+DAUNO+E	ARAC (1); auto-transplant			D
UD4	caratteris	H	Cytogenetic and Molecular	FLAI	ARAC + IDA (2)			D
UD5	caratteris	H	NA	FLAI + IDA ARAC	HD ARAC (1); auto-transplant			CR
UD6	M2	NA	Molecular	FLAI	ARAC+IDA(2); LD ARAC(10)			D
UD8	M0-M1	S	Molecular	FLAI	ARAC+IDA; HD ARAC	auto-transplant		CR
UD9	M0	I	Molecular	FLAI	ARAC(1); IDA+ HD ARAC(1)	LD ARAC		CR
UD10	M4	S	Molecular	FLAI	ARAC+IDA(1); HD ARAC(1)	HD ARAC		D
UD11	M5	H	Molecular	FLAI	ARAC+IDA(1); HD ARAC(1)	CHT + ARAC	allo-transplant	D
UD12	M1	H	Cytogenetic and Molecular	FLAI	ARAC+IDA(1); HD ARAC(1)	allo-transplant (sibling)		D
UD13	M1	I	Molecular	FLAI	ARAC+IDA(1); HD ARAC(1)	MEC (2)	allo-transplant	CR
UD14	M4	S	Cytogenetic and Molecular	FLAI + IDA ARAC	ARAC+IDA(1); HD ARAC(1)	auto-transplant	allo-transplant	D
UD15	M1	S	Molecular	ARAC+DAUNO	FLAI (4); HD ARAC			D

For the majority of the patients, the clinicians assessed the genetic status of FLT3 and NPM1, the two main genetic variants and more commonly mutated genes in AML. As discussed in the Introduction (paragraph 1.3.1) these two types of mutations are important markers for disease predictions. Therefore, the mutational status of FLT3 and NPM1 was used by the clinicians, not only to characterize the samples at diagnosis, but also after treatment, to assess the molecular remission of the disease. In our cohort, this information is available for all patients except for NPM1 in 1 patient at diagnosis, and only for a few patients in the remission and relapse phases of the disease (Table 4.11). In the primary tumours 2 patients had a mutation in FLT3 (one of them had an elevated risk), 7 had a mutation in NPM1 (4 had a standard risk and 3 high risk), other 2 presented both mutations together and they were classified as high risk.

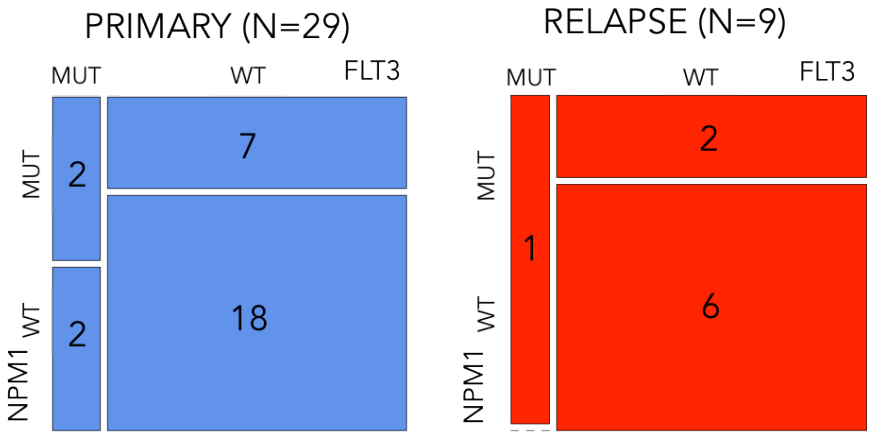
Table 4.11: The mutational status of FLT3 and NPM1 in the three phases of the disease of our cohort of patients. The mutational status is reported as wild type (WT) if the mutation has been tested but was not detected; mutated (MUT, red cells) if it is present; if the test was not performed we reported the not availability symbol (NA, grey cells).

	PRIMARY		REMISSION		RELAPSE	
PATIENT	FLT3	NPM1	FLT3	NPM1	FLT3	NPM1
BO1	WT	WT	NA	NA	NA	NA
BO2	WT	WT	NA	NA	NA	NA
BO3	MUT	MUT	NA	NA	NA	NA
BO5	WT	WT	NA	NA	NA	NA
BO6	WT	WT	NA	NA	NA	NA
BO7	WT	WT	NA	NA	NA	NA
BO8	WT	WT	NA	NA	NA	NA
BO9	WT	WT	NA	NA	NA	NA
BO17	WT	WT	NA	NA	NA	NA
TO1	WT	WT	NA	NA	NA	NA
TO2	WT	MUT	NA	WT	NA	MUT
TO3	WT	WT	WT	WT	WT	WT
TO4	MUT	WT	WT	WT	NA	NA
TO6	WT	WT	NA	NA	NA	NA
TO7	MUT	WT	NA	NA	NA	NA
TO8	WT	NA	WT	NA	NA	NA
UD1	WT	WT	NA	NA	WT	WT
UD2	WT	WT	NA	NA	WT	WT
UD3	WT	MUT	NA	WT	MUT	MUT
UD4	WT	WT	WT	WT	NA	NA
UD5	WT	MUT	NA	WT	NA	NA
UD6	WT	WT	NA	NA	NA	NA
UD8	WT	MUT	WT	WT	WT	MUT
UD9	WT	WT	NA	NA	WT	WT
UD10	WT	WT	WT	WT	WT	WT
UD11	MUT	MUT	WT	WT	NA	NA
UD12	WT	MUT	WT	WT	WT	NA
UD13	WT	WT	WT	WT	WT	WT
UD14	WT	MUT	NA	WT	WT	MUT
UD15	WT	MUT	NA	WT	NA	MUT

Of the 10 patients analysed for FLT3 ITD both in the primary and in the relapse samples, 9 maintained their wild type status in the relapse and one gained the ITD in the relapse sample. The mutational status of NPM1 in the relapse, instead, is always identical to the mutational status detected in the primary tumours: 5 cases remains mutated and 6 remains WT. Furthermore, the co-occurrence and

mutual exclusivity between the two mutations looks very similar both in the primary and the relapse samples, though the numerosity of the patients between the two groups is very different (Figure 4.29).

Figure 4.29: The combinations of FLT3 and NPM1 mutations in the primary and relapse tumours is very similar. The number of patients in each group is very different because mutations are tested more rarely in the relapse samples than in the primary tumours. Despite the different numerosity, the landscape of combinations of the mutations looks very similar in the two groups.

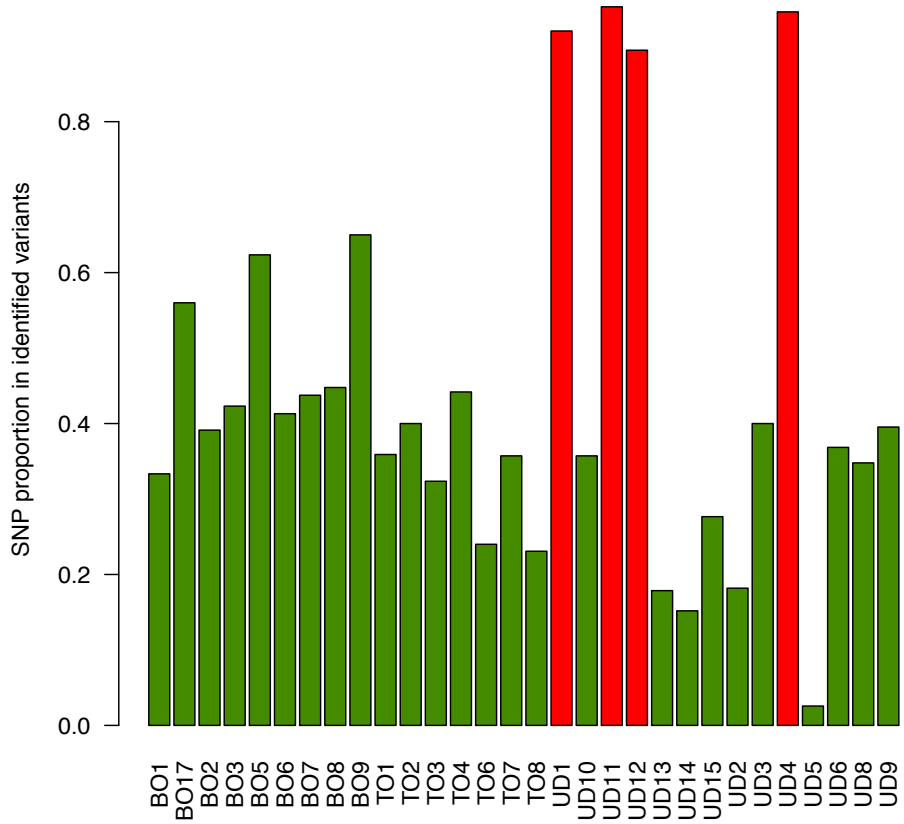


4.2.2 We subtracted the donor variants from the relapse samples obtained after allogeneic bone marrow transplant

We noticed that four samples (UD1, UD4, UD11 and UD12) presented an aberrantly high number of mutations (median number of raw mutations *per* patient 15'635.5), considering all the mutations found in the primary and relapse samples of our cohort (30 patients, median of raw mutations *per* patient in the remaining 26 was 249). Moreover, we noticed a great disproportion in the number of SNPs compared to novel mutations (as shown in Figure 4.30) for these patients: they presented a rate of SNPs over 80%. We tested by Illumina MiSeq 985 SNPs in the highly mutated samples and we validated 943 of them (96%).

Only 4 SNPs were present also in the primary sample of the same patients at frequencies between 1 and 38%. The validation of these mutations revealed that we were not witnessing a sequencing artefact.

Figure 4.30: Proportion of variants overlapping with dbSNP. For every patient we divided the number of variants present in the dbSNP database for the total number of variants identified (in primary and relapse samples). 4 patients had almost 100% of the variants overlapping with SNPs (red bars).



Very often AML patients undergo allogeneic bone marrow transplantation and, of course, this causes a contamination of the relapse leukaemias by cells coming from the donor that repopulated the bone marrow. In the four hypermutated samples, allogeneic transplantation preceded bone marrow collection at relapse, thus resulting in the presence of donor cells in the relapse. Since the DNA of the 4 donors was available, instead of discarding these 4 contaminated samples from our dataset, we decided to proceed experimentally by sequencing also the donor

DNA and subtracting the SNPs identified for each donor from the corresponding contaminated relapse sample. After this subtraction, the number of mutations for the four samples dramatically decreased together with the fraction of SNPs in the samples. We identified a total of 47 mutations (28% already registered as SNPs) for UD1, 8 mutations (12% SNPs) for UD4, 19 mutations (68% SNPs) for UD11 and 33 mutations (15% SNPs) for UD12. These are numbers of mutations and also SNP rates comparable to the ones observed in all other patients of our cohort (median of refined mutations *per* patient 39.5).

4.2.3 The majority of SNVs and Indels are private for primary or relapse tumours. Common mutations affect mostly "landscaping" genes

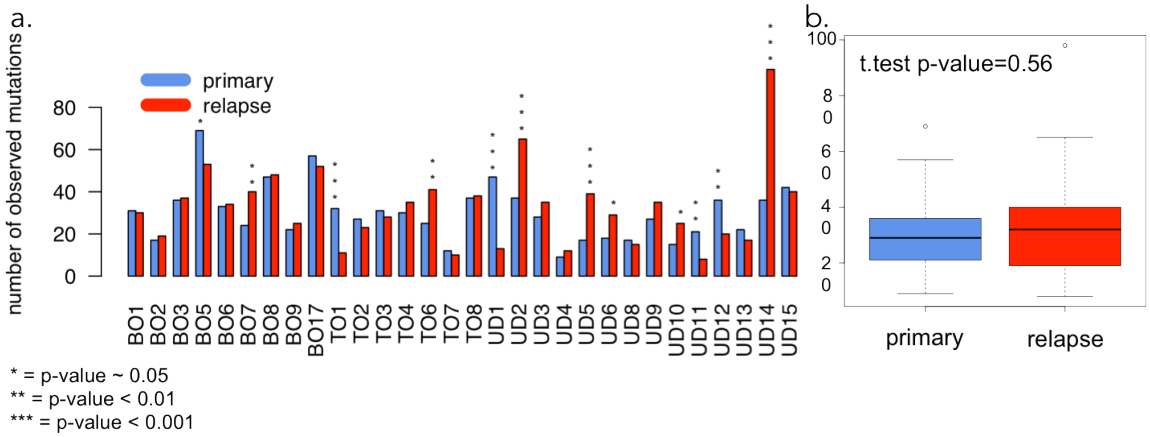
We started analysing the landscape of SNVs and small indels for the primary tumour and relapse samples of each patient, using MuTect and Pindel to identify these variants, respectively. In order to contextualize our results in the AML mutational landscape described in the Introduction section we characterized the variants identified for their known role in the development of the disease and their biological function.

4.2.3.1 The number of primary and relapse specific mutations are similar, but the type of mutations are not the same

The analysis of exomic variants and small insertions/deletions of the primary and

the relapse tumour pairs show significantly different results in many patients, although the median number of variants (SNVs + Indels) per patient is very similar: 29 for the primary tumour (range: 9-69, average: 30.1) and 32 for the relapse (range: 8-98, average: 32.5). Indeed, the p-value of the t-test used to measure the equality of the means of the number of mutations in the two groups is not significant, suggesting that the two populations behave very similarly in terms of number of mutations or, alternatively, the two groups may be too small to exhibit their divergence (Figure 4.31). Nonetheless, 40% of the patients (12/30) show significant differences in the number of mutations detected in the primary *vs* the relapse samples. Such differences not always tend in the same direction: 5 cases show significantly higher number of mutations in the primary tumour and 7 in the relapse (Fig 4.31).

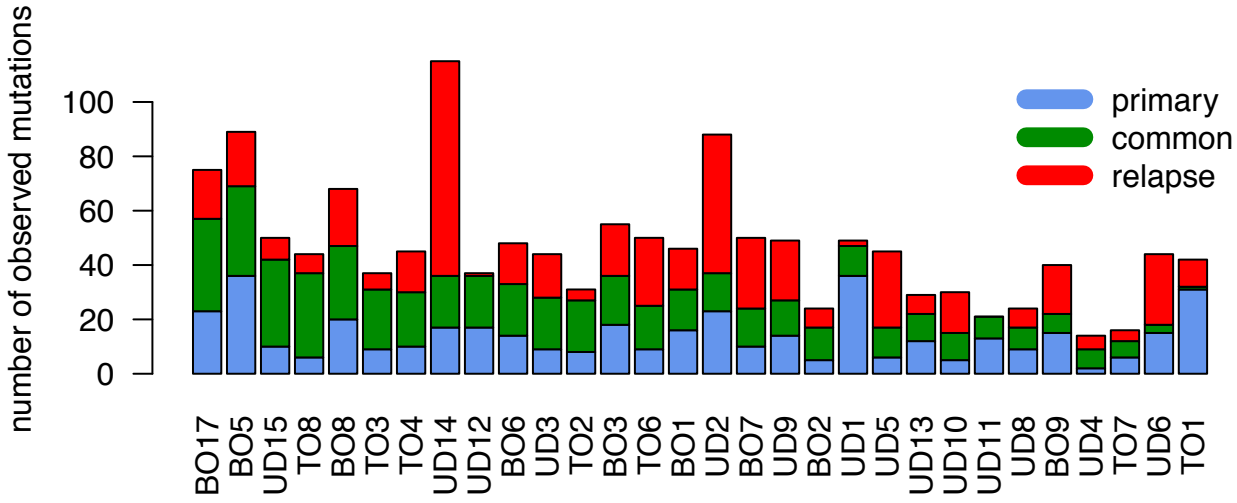
Figure 4.31: The number of mutations detected per patient in 30 AML samples. a. For every patient is reported the number of SNVs and indels identified in the primary tumour (blue bar) and in the relapse (red bar). The stars indicate the p-values (see the legend) obtained testing the difference in the proportion of mutations in the primary and relapse with a statistical test on the proportions. b. The boxplot shows the distribution of the number of mutations in the two groups of samples. Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



Having a similar number of mutations does not necessarily imply that the

mutations are the same in the primary and relapse tumours. As a matter of fact, on average, only 54% of both the primary mutations (range 3-84%) and the relapse mutations (range 9-100%) are in common between the two groups and the remaining 46% are specific for one of the two samples (Figure 4.32). These results indicate that about half of the mutations disappear after chemotherapy and many others are generated. In particular, primary specific mutations were effectively killed by the treatment; at the same time we always observe in the relapse the withdrawal of a group of mutations present in the primary tumour, fostering the hypothesis that chemotherapy might overwhelm at least a part of the tumour population. Emblematic cases are: UD11 with no new mutations in the relapse; TO1 and UD6 with very few mutations (respectively 4 and 16%) in common between primary and relapse (Figure 4.32).

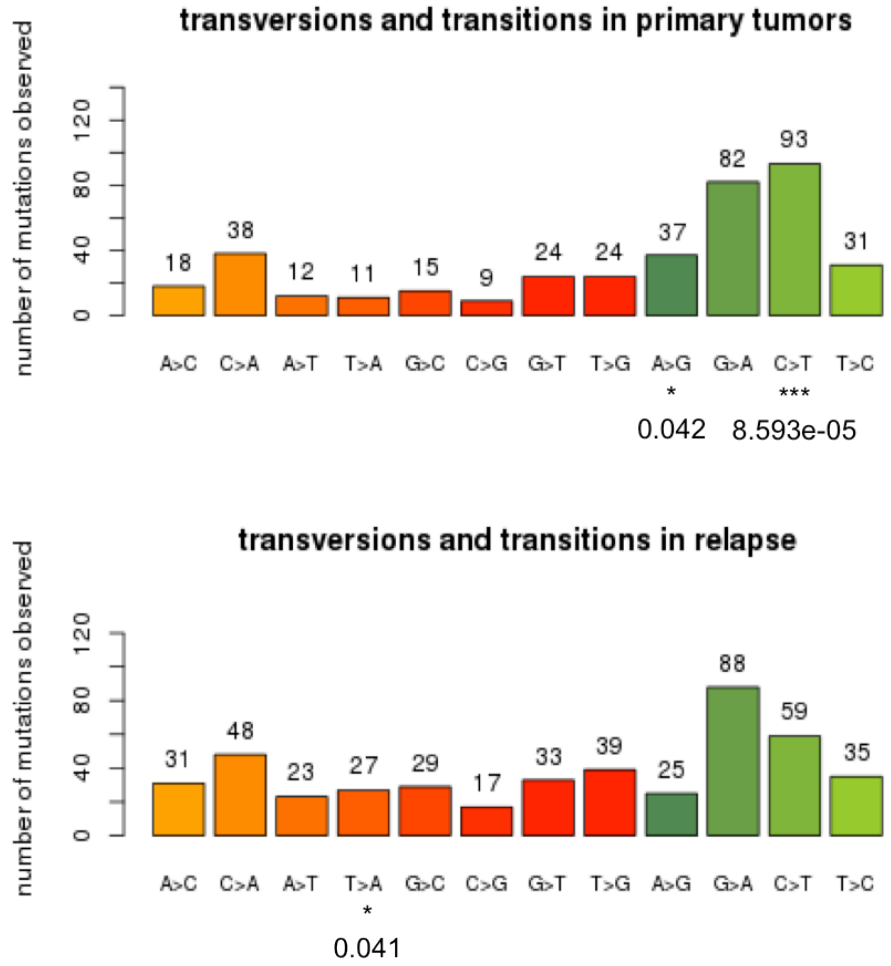
Figure 4.32: Proportion of mutations unique or in common between primary and relapse samples. A stacked bar plot is reported for every patient with the proportion of mutations unique for one of the two samples (blue and red for primary and relapse tumours, respectively) or in common (green). The patients are ordered by decreasing number of common mutation in order to group patients with similar characteristics.



As described in the Introduction (section 1.8), the reciprocal proportion of transitions and transversions can be associated to the mechanism that induced

such mutation types (Ding et al.[133]). Therefore, we inspected their numbers in our AML samples and, through a statistical test on equality of proportions; we tried to identify whether any particular type or more types of mutations were more frequent among the primary and the relapse samples. We found significantly more transitions in the primary tumours (in particular the A>G and C>T type of mutation) and more transversions in the relapse (T>A), confirming previous observations (Figure 4.33). While transitions are more common and they naturally occur in the genome, transversions have been associated to chemotherapeutic agents or other mutagens. Therefore, an augmented proportion of transversions in the relapse samples suggests that some chemotherapy-induced mutations have been fixed in the relapsing cells.

Figure 4.33: Mutations found in the primary tumours (top) and relapse tumours (bottom) gathered by base change. Colours in the red range are associated to transversions, colours in the green range are associated to transitions. If the results show significant differences between primary and relapse samples a star with the p-value is reported.



4.2.3.2 Mutations in AML driver genes often persist after chemotherapy

To understand the role of known AML driver mutations (as defined in Materials and Method, section 3.8), we investigated their presence in the primary and relapse samples and correspondent VAFs. We were able to divide the AML driver genes mutated in our cohort in 7 classes on the basis of their evolution in the samples (Figure 4.34):

1. Genes that always persist: they are found always both in the primary and the relapse samples. These drivers probably resist to chemotherapy and lead to relapse expansion. DNMT3A, IDH2 and EZH2 are "landscaping"[61]

genes that belong to this class; their VAFs are similar or decreasing in the relapse samples, however, the difference in sample cellularity can be the reason for such a reduction;

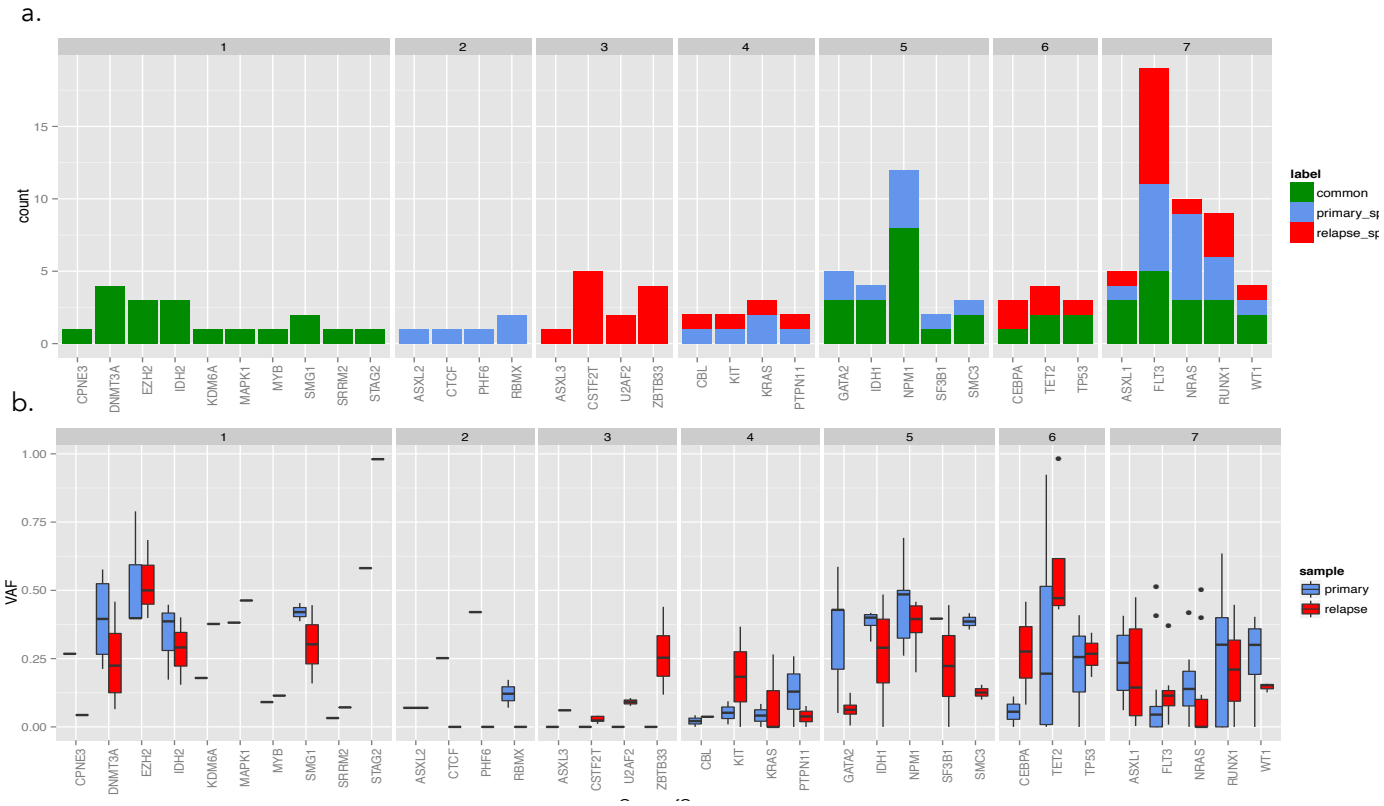
2. Genes that are found only in the primary tumour: these genes are always killed by chemotherapy, thus their presence could make the cells prone to respond to the treatment. This category includes 4 genes implicated in transcriptional regulation: PHF6 harbouring a zinc-finger domain, CTCF that binds chromatin, ASXL2 that is a putative polycomb protein and RBMX, which binds RNA regulating the processes that take place before and after transcription;
3. Genes that are found only in the relapse: since we are looking at a group of AML driver genes, it is probable that they have been found in the primary tumour of other patients. Nevertheless, it is possible that they arise preferentially in a context of tumour predisposition or as a consequence of other mutations: certainly we are looking at single mutations but the mutational landscape underling the appearance of new mutations can already be pre-leukemic. Genes belonging to this group promote RNA polyadenilation (CSTF2T), splicing (U2AF2) and chromatin remodelling (ZBTB33);
4. Genes that are never in common: despite these genes are sensitive to the therapy, probably they are important for tumour development and their presence is needed for a frank leukaemia. In this class, we found well known oncogenes as CBL, KIT, KRAS and PTPN11 that act in the signalling

pathway that promotes a proliferative advantage and that, at the same time, are fundamental for the oncogenic phenotype and make the cells more susceptible to the treatment;

5. Genes that never appear newly in the relapse (see Figure 4.34): these genes are always common or primary only; the reasons for the absence of new mutations in the relapse samples for these genes are ambiguous and it is possible that on a larger cohort we could have observed also their emergence in the relapse samples. The relevant point is the fact that they are recurrently common (the numbers are higher than the previous groups) thus suggesting a possible cooperating role in relapsing tumours. This is a functionally miscellaneous class ranging from transcriptional (GATA2) and cell proliferation activating genes (NPM1), to genes that belong to the spliceosome (SF3B1) and the cohesin complex (SMC3);
6. Genes that persist or appear in the relapse: these genes are always common or relapse specific and have characteristics more adherent to the "resistant" phenotype because when they are present in the primary tumour, at least in our cohort, they resist to therapy; at the same time they can also appear newly in the relapse. Furthermore, the VAFs associated to this class grow in the relapse sample describing an expansion of the relative clone. Interestingly CEBPA, which is generally associated to a favourable outcome, falls in this category, together with TET2 and TP53 both described in the introduction for their known oncogenic role.
7. Genes found in all categories: probably because they are mutated in many

patients, we could observe all the possible combinations. These are all well known leukaemia-associated genes and it is possible that their cooperation with other genes of the same or other classes contribute to therapy resistance.

Figure 4.34: AML driver genes often persist after chemotherapy. a. For every AML driver gene mutated in our samples we report how many times it is uniquely mutated in the primary, in the relapse tumour or in common. We then divided the genes on the basis of their behaviour in this analysis. b. For every AML driver gene, the boxplot shows the VAFs identified in the primary (blue) and relapse tumours (red). Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



4.2.3.3 DNA methylation and Cohesin complex mutations persist in the relapse, spliceosome mutations disappear after chemotherapy

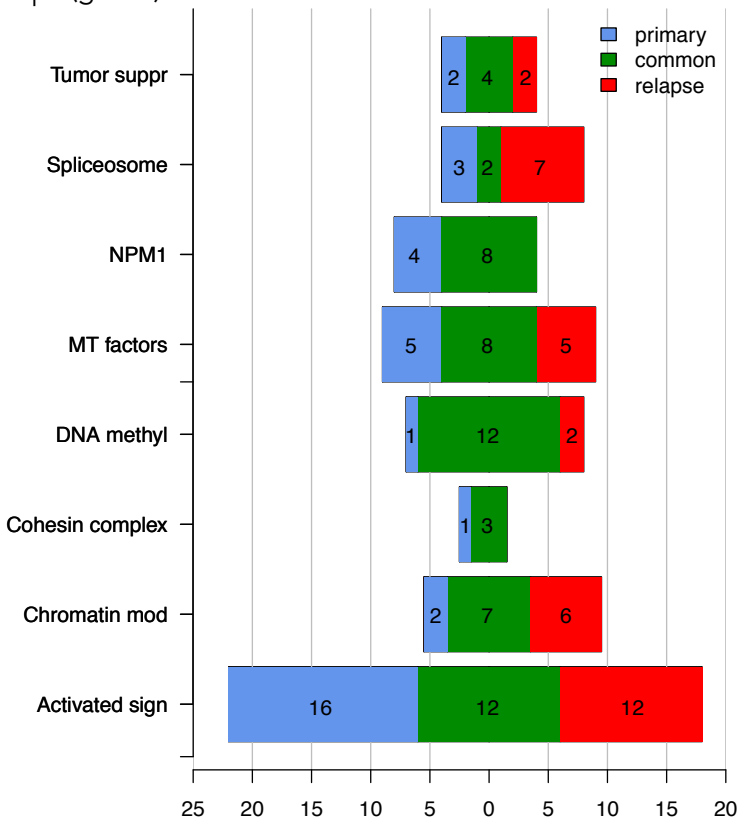
We then divided the mutated genes into functional categories according to the paper of the TCGA[44]. We observed three general trends (Figure 4.35):

- Categories of mutated genes that persist in the relapse: it has been,

previously, hypothesized that "landscaping"[61] mutations would escape chemotherapy. In fact, we identified the functional class of genes involved in DNA methylation common to the majority of cases. Also mutations that never occur in the relapse belong to this class, they can be ascribed to a very small cohort and they are frequently common between primary and relapse, suggesting that this classes of genes are able to survive the treatment (NPM1, cohesin complex);

- Categories that are mostly not persistent: variants affecting the spliceosome machinery seem to be more susceptible to chemotherapy, however, they often appear in the relapse tumour with novel mutations;
- Categories with no prevalent behaviour: in this class fall Tumour suppressors, Myeloid transcription factors, chromatin modifiers and Activated signalling genes; their heterogeneous presence in cells respondent or not to therapy together with their concomitant new emergence in relapse cells suggests that they either were already present in a minor fraction of cells at diagnosis that expanded after treatment or that they need the co-occurrence of other mutations to impart a phenotype.

Figure 4.35: DNA methylation and Cohesin complex mutations persist in the relapse, spliceosome mutations disappear after chemotherapy. For every functional category described by the TCGA, the graph reports the number of mutated genes identified uniquely in the primary (blue) or in the relapse tumours (red) or in common in the two groups (green).



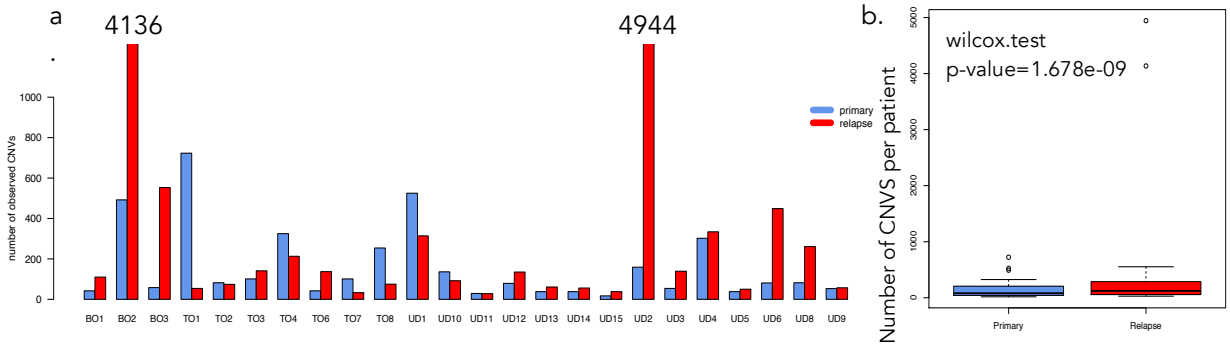
4.2.4 Common CNVs are very rare and poorly defined from a functional point of view

Copy number variants were analysed in our cohort of patients using the Control-FREEC tool on the WES data. We were able to perform this analysis only on 24 patients, because the remaining 6 were sequenced after exome enrichment with two different kits for primary and relapse tumours. This resulted in an aberrantly high proportion of CNVs, very likely due to different coverage of the exonic regions of the genome by the two kits.

4.2.4.1 The CNVs are very variable among patients and samples and they are seldom retained

We identified a median number of CNVs per patient of 81.5 (mean: 160.5) for the primary tumours and 122.5 (mean: 520.2) for the relapses, with ranges from 17 to 723 and from 28 to 4944, respectively (Figure 4.36). In particular, two patients (BO2 and UD2) behave as outliers concerning the number of the CNVs in the relapse, which presents more than four thousands of variants each. These AMLs display a number of CNVs above average also in the primary tumour, suggesting that a complex genomic status in the primary tumour could degenerate in a catastrophic genomic event under the pressure of chemotherapy. In general, our patients revealed a great variability in terms of number of CNVs. All patients, with the exception of 5 patients (TO2, UD11, UD4, UD5 and UD9), have significant different number of CNVs when tested for the difference of proportions between variants identified in the primary tumour *vs* the relapse (Figure 4.36). Because of the great variability observed between primary and relapse tumours, we can detect a significant preponderance of CNVs accumulation in the relapse tumours compared to primary tumour within our cohort (p-value 1.678×10^{-9}). Furthermore, no correlation has been identified between the number of nucleotide variants and copy number variants neither sample wise (ρ=0.013), nor patient wise (ρ=-0.14) (data not shown).

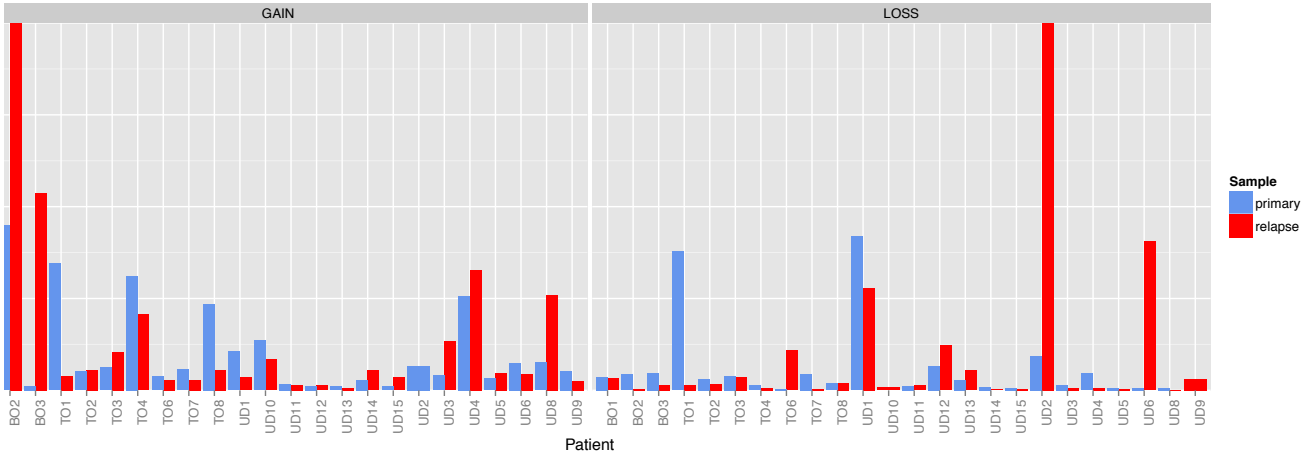
Figure 4.36: The variability in copy number abundance among patients is high. a. For every patient, we show the number of CNVs identified by Control-FREEC both in the primary (blue bars) and in the relapse tumour (red bars). b. We reported the boxplots of the number of CONVs identified per patient in the primary and relapse samples. The p-value for the Wilcoxon-test comparing the means in the two populations is reported (Shapiro test was used to assess the normality within groups, resulting in a significant difference from the normal distribution both for the primary and relapse samples). Boxes define the 25th and the 75th percentiles; horizontal line within the boxes indicates the median and whiskers define the 10th and the 90th percentiles.



We, next, categorized the aberrations of CNVs in losses and gains based on the number of copies present in the tumour compared to the control: losses if the tumour has fewer copies than the control; gains if the tumour has more copies than the control. In general, we observe slightly more gains than losses (188 gains vs 163 losses per patient, on average) (Figure 4.37). However, there is no statistical difference between the means of CN gains and losses between the primary and the relapse tumours considering all patients (p-values 0.35 and 0.36, respectively, with confidence intervals for the difference of the means of [-543,198] and [-619,232]) and there is no statistical difference in the numbers of gains and losses between primary and relapse samples of each patient (Figure 4.37). Finally, the CNVs detected in the two samples with an impressively high number of variants may be false positives. Indeed, all these CNVs are scored almost exclusively as gains for BO2 and as losses for UD2 (Figure 4.37). This may

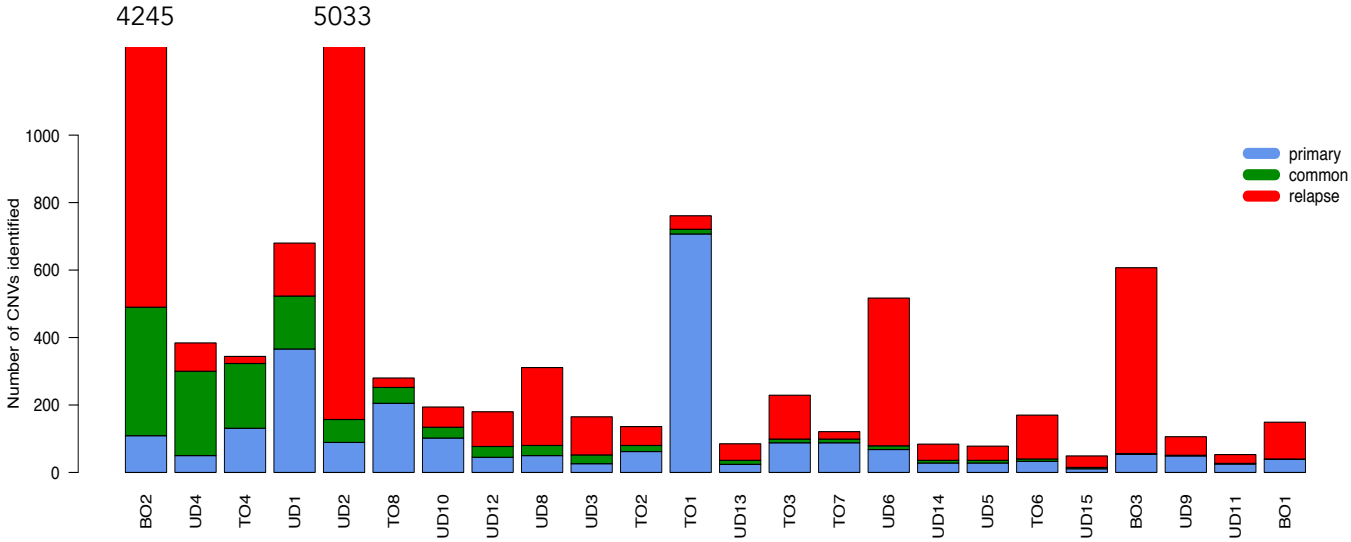
be imputed to a misinterpretation of the data by the algorithm likely due to enrichment bias.

Figure 4.37: In our cohort of patient there is no preponderance of CN losses or gains. We segregate CNVs in gains (GAIN) and losses (LOSS). The bars represent the number of CNVs detected in the primary (blue) and in the relapse tumours (red).



In parallel with the analysis performed for the nucleotide variants, we wanted to understand how many of the variants scored in the primary tumours are maintained in the remission samples of the same patient. The proportion of CNVs retained after chemotherapy is quite variable but, in general, very low: in respect to the primary tumour on average 27% (ranging from 2 to 83%), the 42% of the primary tumours (10/24) has less than 20% of common mutations; in respect to the relapse they are on average 22% (ranging from 0.4 to 90%; Figure 4.38). Indeed, the vast majority (median 85%) of CNVs detected in the relapse are new (Figure 4.38).

Figure 4.38: A quite small proportion of CNVs detected in the primary tumour is retained in the relapse. For every patient, we report the number of CNVs identified uniquely in the primary (blue) or the relapse (red) or in common to both (green).



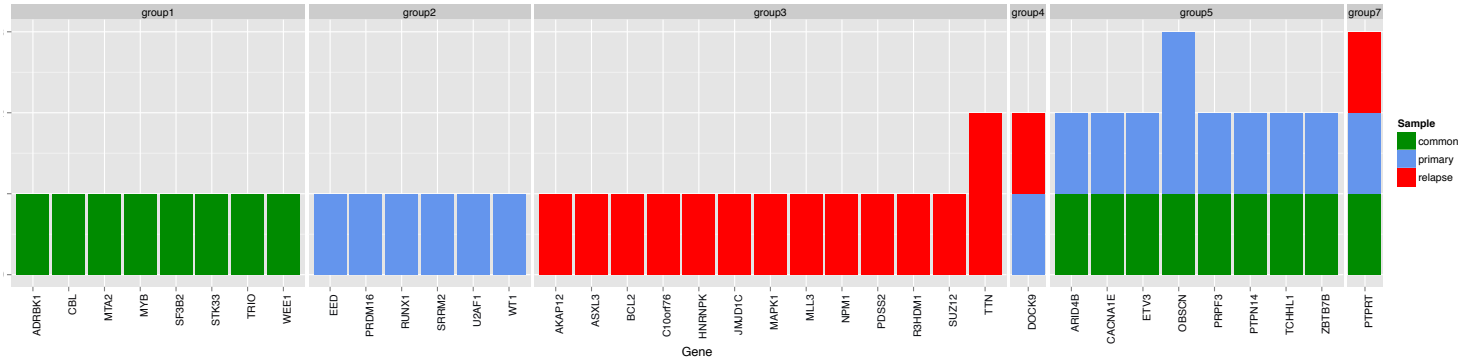
4.2.4.2 The majority of driver genes involved in CNVs are not recurrent

We then asked how many and which AML driver genes are affected by CNV gains or losses. From this point on, we decided to eliminate from our analysis the two patients (BO2 and UD2) that probably presented many false positives, in order to eliminate possible bias introduced in the analysis by the high number of gains and losses identified in either one of them. We detected 61 driver genes affected by CNVs, the majority of them (61%) were hit by only one CN gain or loss in our cohort. Only two genes were hit at least 3 times in our cohort: PTPRT, a signalling protein involved in cell growth and differentiation, with 3 gains and one loss, and EZH2, a transcriptional repressor, with 3 losses (Figure 4.39 and Figure 4.40); in both cases the CNV was found both in common and specifically in the primary or the relapse tumour of different patients. We performed our analysis on CNVs,

considering separately gains and losses.

As shown previously for SNVs, we divided the results in 7 groups based on the behaviour of the gains affecting the AML drivers in the different tumour phases (Figure 4.39). Genes belonging to a specific group, in this case, appear to have more mixed functions as compared to the results shown before for SNVs. Many of the genes mutated by CNVs that belong to Group1 and, therefore, maintained in common among primary and relapse tumours, are serine/threonine kinases (STK33, WEE1), GTPases (ADRBK1, TRIO) and oncogenes (CBL, MYB). Group5 genes, in common or primary-specific, are mostly related to regulation of transcription (ARID1B, ETV3, PRPF3) and calcium binding (CACNA1E, TCHHL1).

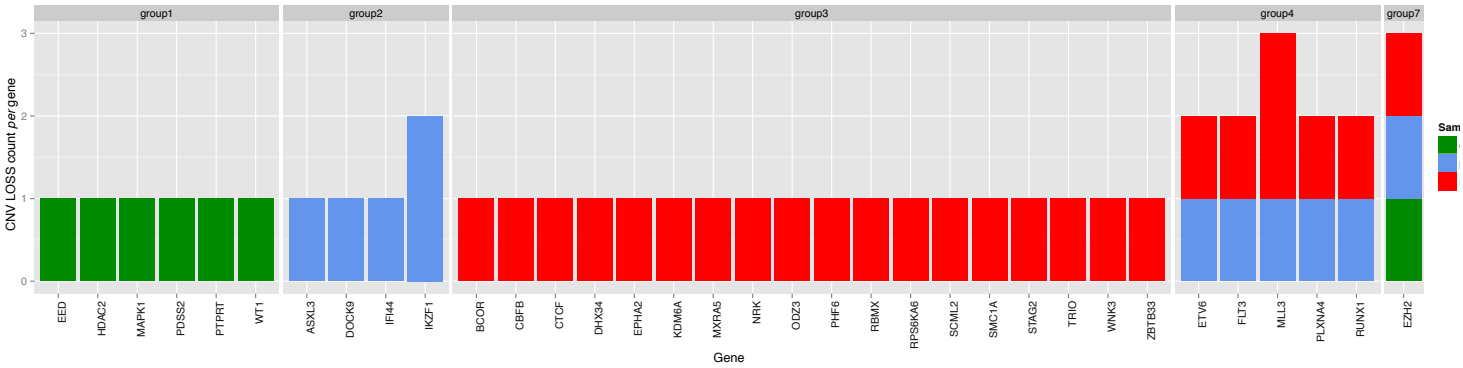
Figure 4.39: AML driver genes hit by CN gains in our cohort of patients. We report the number of times (CNV GAIN count *per* gene) each AML driver gene was hit by a CNV gain. The AML drivers were divided into groups as described for SNVs in Figure 3.32. The bars show the number of gains identified in each gene uniquely in the primary (blue), in the relapse tumours (red) or in common between the two groups (green).



The same type of analysis on CV losses reveals that CNVs retained after chemotherapy (Group1 and Group7) include a tumour suppressors (WT1), landscaping mutations (HDAC2, EED interacting with EZH2), kinases (MAPK1) and phosphatases (PTPRT). Genes belonging to Group2 and Group4, never in common between primary and relapse samples of the same patient or primary specific, include many factors that regulate transcription (ASXL3, ETV6, IKZF1,

RUNX1) and signalling genes (DOCK9, FLT3 and PLXNA4) (Figure 4.40).

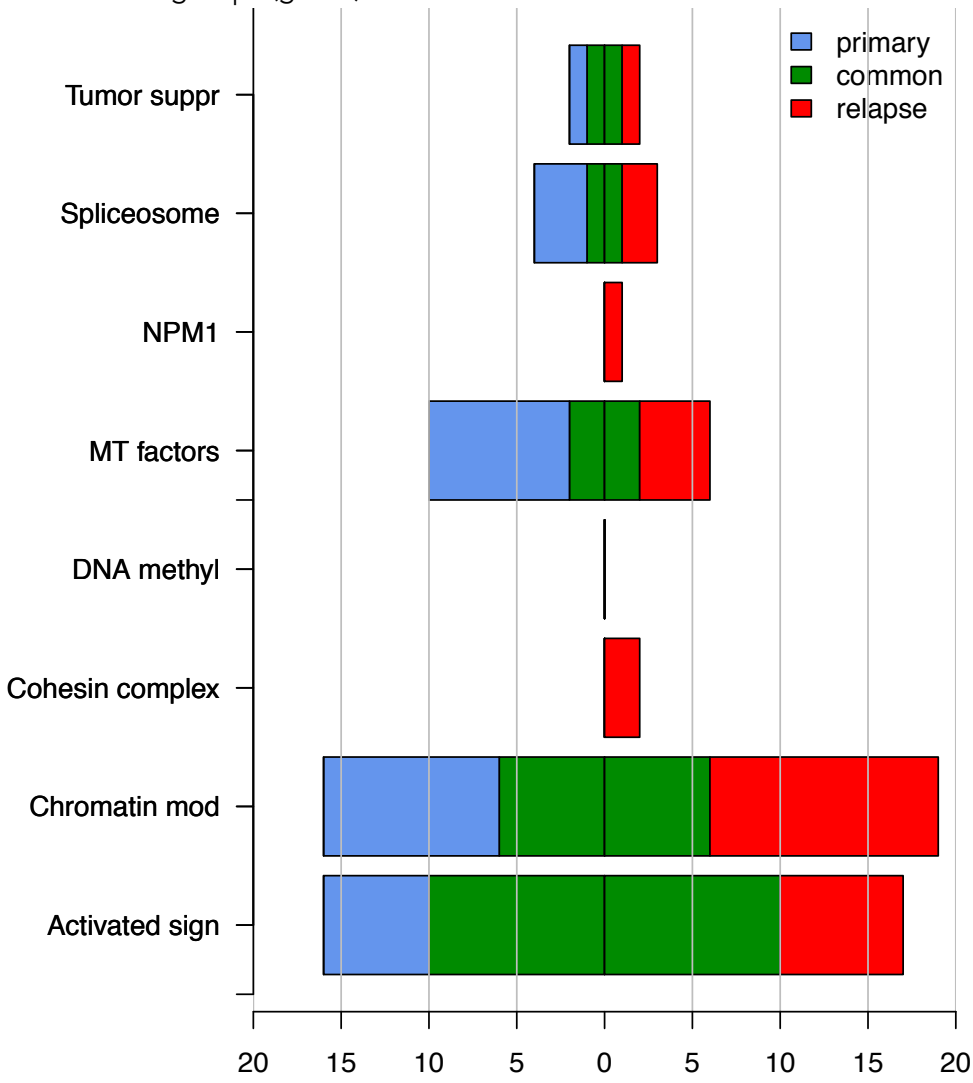
Figure 4.40: AML driver genes hit by CN losses in our cohort of patients. In the plot is reported the number of times each AML driver gene falls in a region of CNV loss. The AML drivers were divided into groups as described for SNVs in Figure 3.32. The bars show the number of losses identified in each gene uniquely in the primary (blue), in the relapse tumours (red) or in common between the two groups (green).



4.2.4.3 CNVs hitting AML drivers belonging to Activated signalling and chromatin modifiers functional classes are retained in the relapse

Considering the classification of the AML drivers hit by CNVs into functional categories, as described above (paragraph 1.3.2.2) and in [44], it is worth noticing that tumour suppressors, myeloid transcription factors and spliceosome pathways appear rarely in common between primary and relapse, corroborating our previous evidences on mutated AML drivers, showing that transcription factors are often found uniquely associated to one of the two samples (Figure 4.41). CNVs in chromatin modifiers and factors activating signalling pathways can persist after chemotherapy, though the persistent CNVs are not preponderant among all. CNVs in NPM1 and cohesin complex components always appear in the relapse (Figure 4.41).

Figure 4.41: CNVs hitting AML drivers belonging to activated signalling and chromatin modifiers functional classes are retained in the relapse. For every functional category described by the TCGA, the graph reports the number of genes hit by CNVs identified uniquely in the primary (blue), in the relapse tumours (red) or in common between the two groups (green).



4.2.5 Clonal analysis of the tumour populations in our patients' cohort

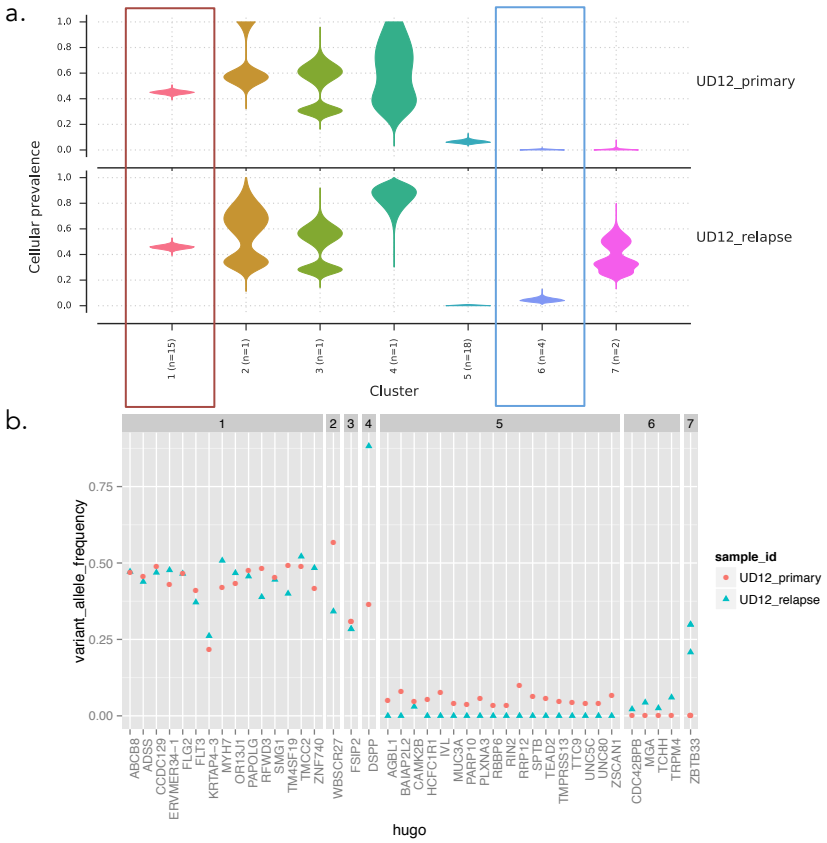
We performed clonal analysis for all 30 patients in our cohort using PyClone. In order to reconstruct the clonal composition of the tumour population, the Pyclone

algorithm requires as input the following parameters derived by WES analysis: i) the variants (SNVs + Indels) identified during the mutation calling phase; ii) the relative frequencies (VAFs) of each variant; iii) the number of copies (CNs), which has been set to 2 for all the regions without CNVs. From the analysis of each patient, PyClone generates a figure similar to the one shown in Figure 4.42 panel a, together with the specification of frequency of the clones and the list of the mutations belonging to each clone. As example of the output of the analysis, we show the results obtained for one of our patient, UD12 (Figure 4.42). For each patient, PyClone produces always two plots: the first is related to the primary tumour clones and the second to the relapse clones. Each clone in the plot is visualized as a "violin", which, for each clone, depicts the probability of having a given cellular frequency in the tumour population. When the number of mutations within a clone is sufficiently high, the violin is small, indicating that the margin of error for the estimate of the frequency is low (clone 1, clone 5 and clone 6 in Figure 4.42.a). In some cases, the clones are represented by only one mutation; for those clones the difficulty to estimate the real cellular frequency is depicted by a large "violin" (clone 2 in Figure 4.42.a). However, if the frequency of the single mutation is quite high and, therefore, the probability of assigning a wrong frequency in the tumour population is low, again the violin is small (clone 4 at relapse in Figure 4.42.a). For every cluster (i.e. clone) identified, we plotted the VAFs of all the mutations assigned to each particular cluster (Figure 4.42.b). First, we eliminated all the clones with a frequency lower than 10% in both samples because the determination of clonal origin at low frequencies is ambiguous and it

is probable that these clones contain new mutations belonging to any clone in the population; then, we classified the clones based on the different frequencies they displayed in the primary tumour *vs* the relapse tumour. Based on their VAFs in the two tumour populations, the clones were named as follows:

- Primary only: if the frequency of the clone in the relapse was lower than 2%;
- Relapse only: if the frequency of the clone in the primary tumour was lower than 2%;
- No change: if the frequency of the clone differed of less than 5 percent points in the two tumour phases;
- Decreasing in relapse: if the difference in frequency was higher than 5 percent points and the clone was more frequent in the primary tumour than in the relapse;
- Growing in relapse: if the difference in frequency was higher than 5 percent points and the clone was more frequent in the relapse than in the primary tumour.

Figure 4.42: PyClone analysis of patient UD12. a. Typical output plot made by PyClone. We highlighted in red an example of clone "not changing" and in blue an example of "relapse only" clone that was, however, eliminated from further analysis because its frequency was lower than 10% in both tumour phases. b. Plot of the frequencies of the mutations occurring in each listed gene in the primary (blue triangles) and the relapse (red dots) tumours grouped by their clonal membership.



Clone 1 "no change" (from 46 to 45% of frequency)

Clone 6 is relapse specific (from ~0 to 5% of frequency)
We will not consider it because its frequency is below 10%

4.2.5.1 The majority of the patients show resistance to chemotherapy

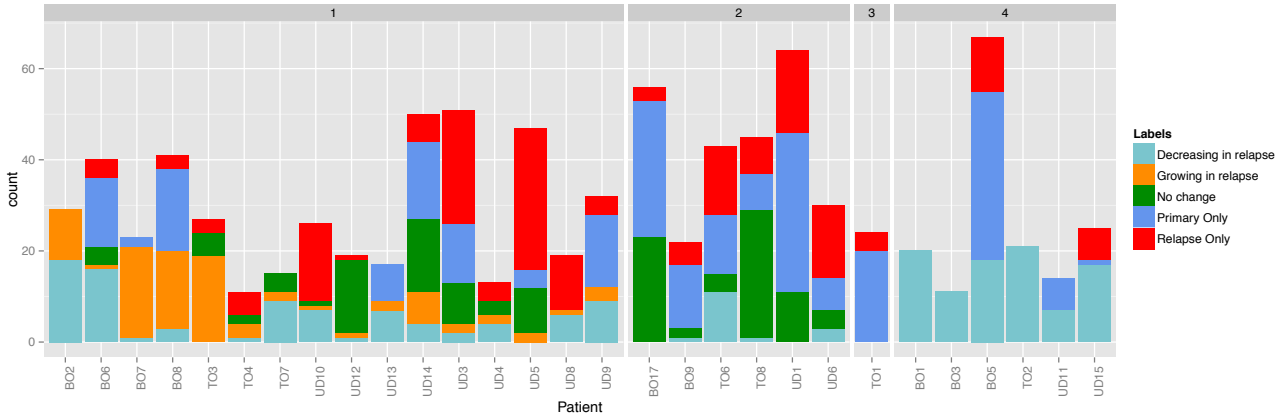
We investigated the clonal composition of all tumours for all patients except UD2, for which PyClone was unable to produce an output. Based on the clonal analysis performed as described for UD12, for every patient, we assigned each mutated gene to one of the classes of clonal membership listed above and we plotted the results in Figure 4.43. We were able to identify 4 possible schemes of tumour evolution after chemotherapy (note that the numbers reported above the boxes in Figure 4.43 refer to the following schemes):

1. CLONAL SELECTION (16, 55% of the patients): belonging to this class are tumours characterized by the presence of a quite high fraction of “no change”, “decreasing in relapse”, “growing in relapse” and “relapse-only” clones, accompanied by a small fraction of “primary only” clones. This clonal composition suggests us the possibility that the therapy affected some clones (those vanishing or diminishing), successively favouring the expansion of more fitted resistant clones that generated the relapse tumour;
2. RESISTANCE (6, 21% of the patients): clones in primary and relapse tumours have similar frequencies (the tumours are mainly composed of “no change” clones), suggesting that a part of the tumour simply resisted to chemotherapy;
3. NO COMMONALITIES (only one patient): primary tumour and relapse have no clones in common (all clones are “primary only” or “relapse only”). Only one tumour, TO1, belongs to this class. The two phases of the disease look as two different tumours with specific and independent clonal populations. It would be difficult to find an explanation for this behaviour, even because the time to relapse for TO1 was of about 12 months, that seems a quite short time to develop a new and separate leukaemia. Indeed, inspecting the VAFs of the variants manually, we are able to score the presence of one cluster with VAFs close to 0%, but not null, in the primary tumour, that slightly increases in frequency in the relapse, however never reaching the 10% threshold. So the result we obtain is due to

analysis thresholds, that are arbitrary, but we still observe a primary clone expanding at relapse;

4. DECREASE OF COMMON CLONES (6, 21% of the patients): all clones present in the primary tumour decrease or disappear (the tumours are mainly composed of "relapse only" and "decreasing in relapse" clones). Even if we can not exclude that there might be some other mechanisms that influence relapse expansion or that the effect of chemotherapy is the introduction of new mutations, leading to relapse formation. This behaviour could partly be an artefact due to a problem of purity of the remission sample, used to call the mutations, impairing the mutation calling task because their presence in the remission challenges the probability to call a somatic mutation.

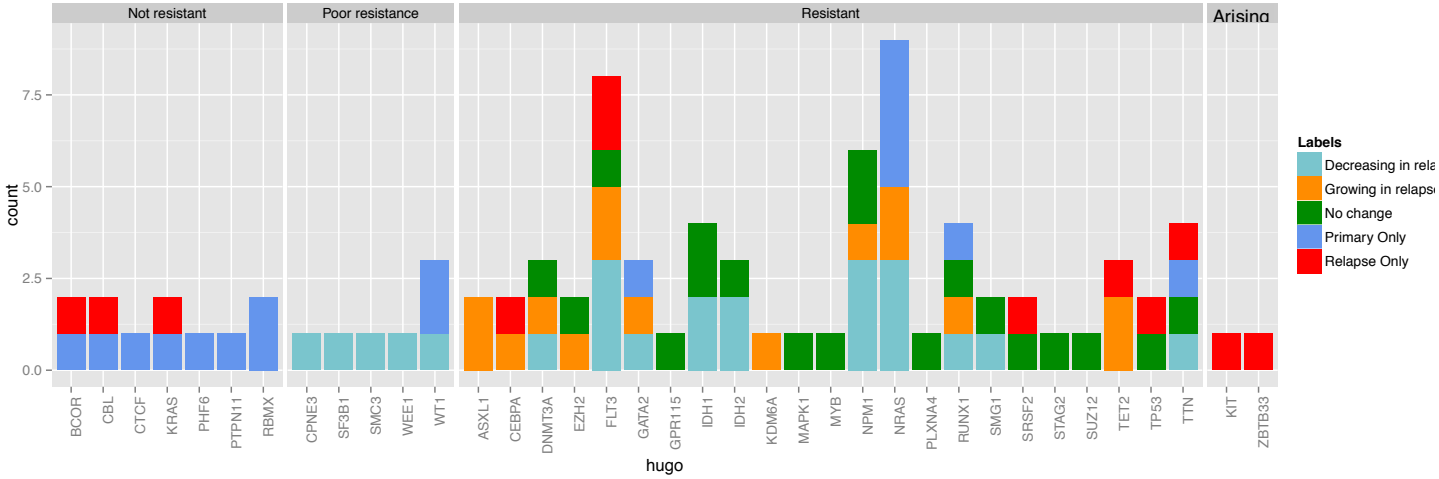
Figure 4.43: Schemes of clonal evolution in our AML samples. For every patient we report the clonal membership of every mutation identified. The colours are referred to the labels assigned to each clone as reported in the Legend and described in the text (paragraph 4.2.5.1). The four boxes correspond to the evolutional behaviour described in the text: 1, Clonal selection; 2, Resistance; 3, No commonalities; 4, Decrease of common clones.



4.2.5.2 Many driver genes resist to chemotherapy

As reported for all genes, we performed a clonal analysis considering only the AML driver genes, grouping them on the basis of the behaviour of the clones that harboured their mutations. They were classified as: "Resistant", if they appear in clones that survived chemotherapy and in some cases expanded in the relapse; "Poorly resistant", when, though they were present at relapse, the frequency of their clones decreased; "Not resistant", if they were present in clones killed by the chemotherapy. Of course, when the mutation was not present or not detectable in the primary tumour, we were unable to categorize them and assigned them to the "Arising" category. "Resistant" drivers are mainly involved in chromatin organization (SUZ12, DNMT3A, KDM6A, EZH2) and positive regulation of biosynthesis (CEBPA, MAPK1, GATA2, NPM1, TP53, RUNX1). Functional annotation of resistant genes attributes them mainly to "landscaping"[61] and cell proliferation pathways (Figure 4.44). The list of AML driver contains some fishy genes like, for example TTN; in our analysis we can recognize the passenger nature of TTN because we observe its presence in clones of many different types, thus suggesting that it is not the primary player for the resistance of those clones (Figure 4.44).

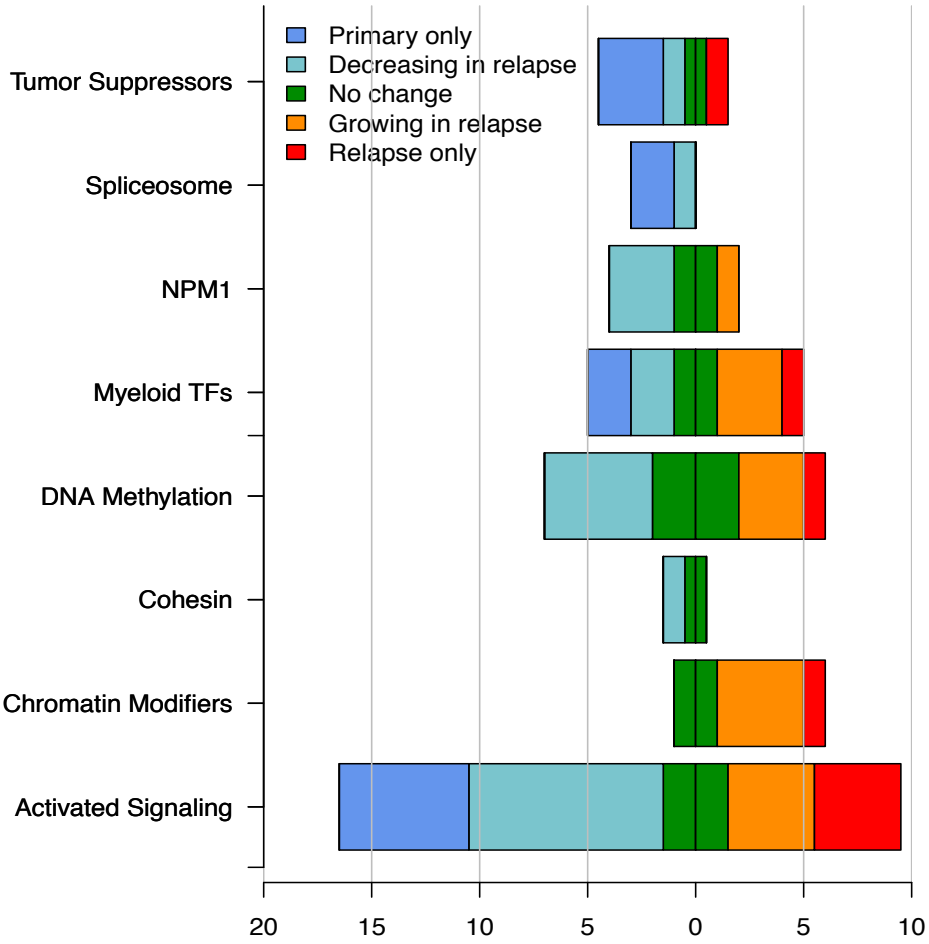
Figure 4.44: Many clones harbouring mutations in AML driver genes are resistant to chemotherapy. Each AML driver gene is assigned to a category based on the behaviour of the clones that harbour their mutations (see text). The same gene in different patients appeared in clones with different characteristics.



4.2.5.3 Clones containing mutations in NPM1 or in genes that belong to chromatin modifiers, cohesin complex, and DNA methylation are resistant to chemotherapy

If we consider, once again, the functional categories to which the driver genes belong (Figure 4.45), based on the classification of clonal segregation, we observe that the spliceosome and tumours suppressor genes functional categories disappear or are reduced after chemotherapy. NPM1, genes involved in DNA methylation and in cohesion complex look mostly unaffected, because they never disappear after treatment. Mutations in chromatin modification pathways reappear at the relapse always at the same or incremented frequency. Mutated genes belonging to activated signalling and myeloid transcription factors have, instead, a more varied behaviour, suggesting that genes in these categories can either be resistant or not, or that they might cooperate with drivers belonging to the same or other categories.

Figure 4.45: Clones containing mutations in NPM1 or in genes that belong to chromatin modifiers, cohesin complex, and DNA methylation are resistant to chemotherapy. For every functional category described by the TCGA, the stacked bar plots represent the number of clones belonging to the categories reported in the figure legend.



4.2.5.4 The remission sample seldom is mutated at low frequency

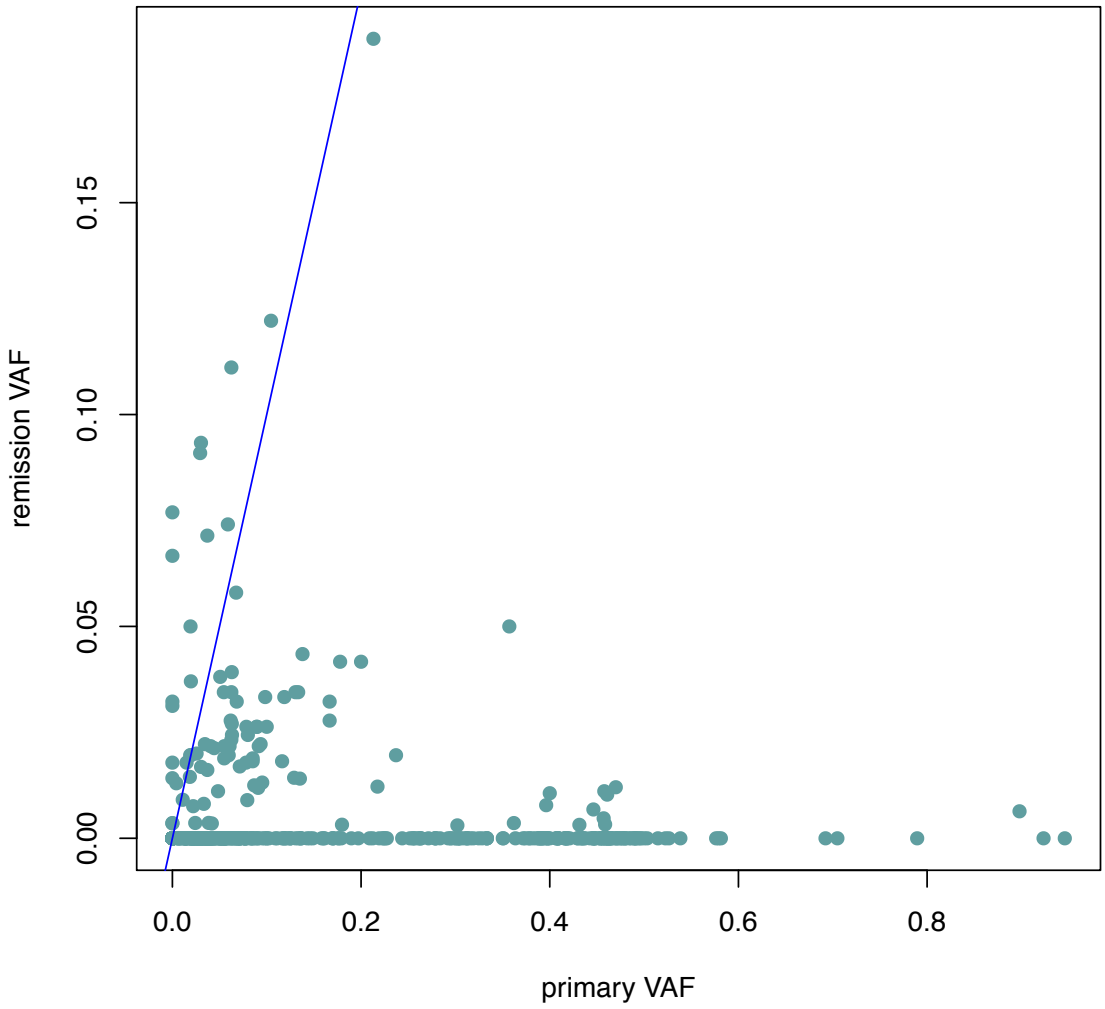
Indeed, the presence of driver mutations at remission has been already shown for AML (see paragraph 1.4); their importance is paramount both from the clinical point of view: they can underlie the presence of minimal residual disease, and for the bioinformatics analysis because they can challenge the variants call. To investigate the presence of mutations at low frequency in the remission sample, we isolated all the genes belonging to the "no change" and "growing in relapse"

classes as we can be confident that the genes belonging to these classes survived chemotherapy. After, we plotted their frequencies in the primary and remission samples; we thought that pathogenic genes would decrease in frequency in the remission and therefore we filtered only genes presenting lower VAF in the remission compared to the primary tumour (Figure 4.46). The final list contained 386 genes. Among them we found many reasonable false positives: all the taste and olfactory receptors, mucins and keratins are families of genes having very similar sequences and high variability among individuals; the alignment in these positions is often tricky (DSPP, TAS2R46, KRT8, MUC2, TCHH, FAM186A, OR1L4, OR2L3, ANKLE1, KIAA0196, KRTAP9-4, OR2L8). At the same time we also found "landscaping"[61] genes (EZH2, ASXL1, IDH1, RUNX1, TET2, SETD8), signalling genes (NPM1, NRAS, SIRPA), genes associated to poor outcome (FLT3, TP53) and to drug metabolism (ABCC12).

In particular, we observed that four genes belonging to clones of the "no change" class, detectable in the remission sample, were AML drivers or were classified as cancer genes by the Cancer Gene Census (CGC). RUNX1 (identified in TO8) had a frequency of 40% in the primary tumour, almost disappeared at remission (1% VAF) and came back at relapse with 30% frequency. A similar behaviour was observed also for CCND2 (in TO8: 40% VAF at primary, 0.8% VAF at remission and 19% VAF at relapse). Finally, other two genes have similar trajectories but at lower VAFs: MLH1 (in TO4) from 4% VAF in the primary goes to 0.4% in the remission and reappears in the relapse at 9%; HSP90AB1 (in BO8) from 5% VAF to 2% and reappears at 7% VAF in the relapse. Note that the VAF of

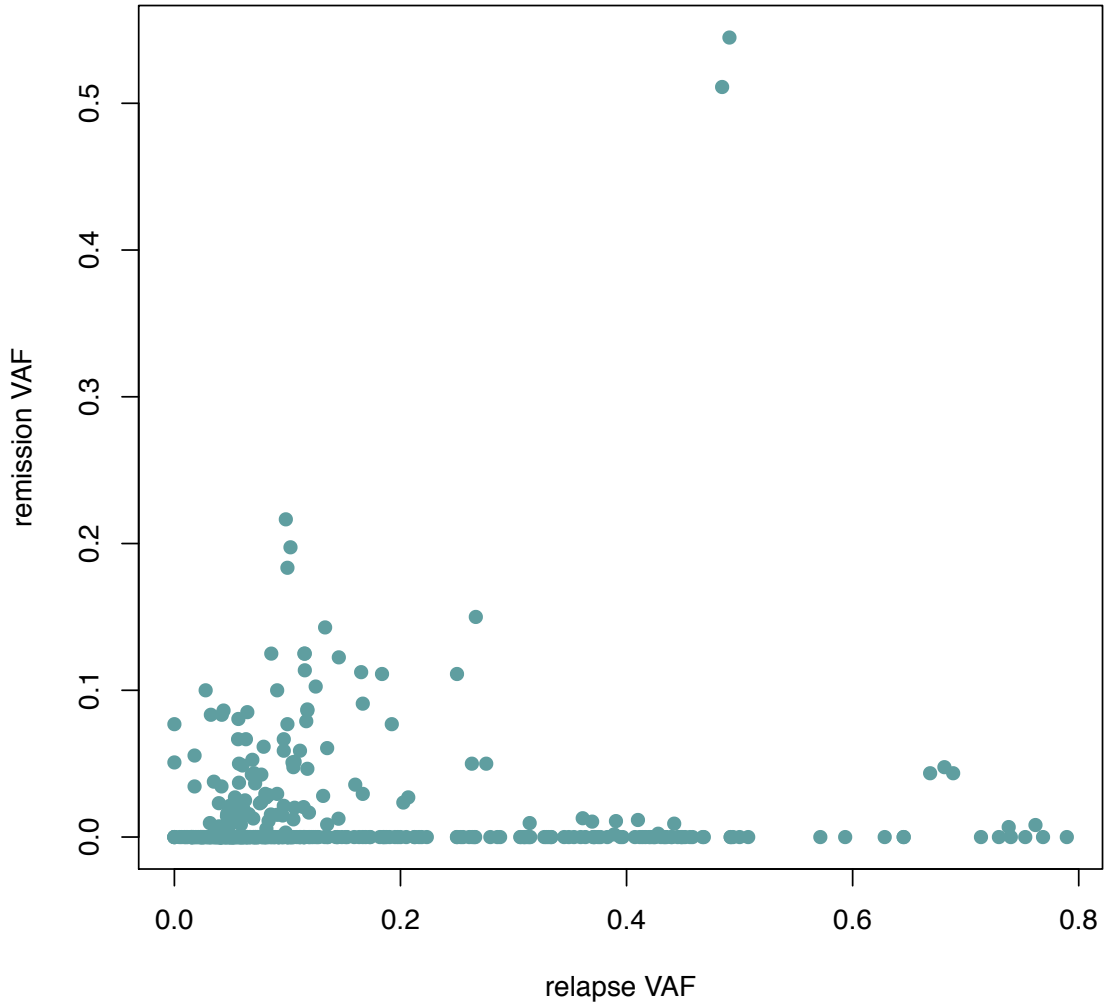
a gene can dissociate from the frequency of the clone it belongs to, therefore, genes belonging to the "no change" class can have VAF differences bigger than 5 percentage points. All these observations underline the problem of remission determination: in the case of TO8, for example, the treatment was very likely not sufficient for the complete eradication of the disease; additional molecular markers for remission assessment would improve the confidence in molecular remission determination.

Figure 4.46: "No change" and "growing in relapse" classes are seldom present in the remission sample. We plot for the mutations belonging to those classes, the VAFs in the primary and remission sample in order to detect their presence in the remission. We also isolated from them the group of mutations having lower frequency in the remission compared to the primary tumour and characterized that list of genes. The blue line bisects the plan, mutations having lower VAF in the remission compared to primary lay below the blue line.



The same analysis was performed on the "Relapse Only" mutation class and we observed that, although the majority of mutations were undetectable at exordium, many of them (119) were already present at remission (Figure 4.47). 22/29 (76%) of the patients had "relapse only" genes detectable at remission; excluding 5 outliers with high VAF of 18%, 20%, 22%, 51% and 54%, at remission they had a median VAF of 3% and a maximum VAF of 15%. Nonetheless, many of them were present on probable drivers, and we could detect AML driver genes, invisible at exordium, mutated at remission only in 4 patients. The genes were U2AF2 (BO6), CRIPAK (BO9 and UD14) and RBMX (UD9). The possible explanations for this event are two: that mutation could be present at very low frequency in the primary tumour and expanded later or chemotherapy induced them. Analysing the primary tumour at deeper sequencing could provide better insights on these events. Since these genes were previously identified as mutated in AML but are not known as primary players in the disease (no NPM1 or FLT3 mutations were found), it would be interesting to determine their presence in a wider cohort of patients. Moreover, we analysed the presence of known drivers from CGC (not strictly AML drivers) in our cohort: patient BO2 presented a "relapse only" mutation detectable at remission on GNAS, a CGC driver gene. UD14, besides CRIPAK (mutated at 5% VAF) had mutations on other driver genes: MN1 (12% VAF), NFKB2 (5.5% VAF) and TSC2 (4% VAF); this can be possibly explained with the presence of a clone that expanded at relapse, becoming ~10% VAF (correspondent to 20% of cells, not the dominant clone).

Figure 4.47: "Relapse only" variants are often detectable at remission. We isolated the variants that were present at relapse but not detectable in the primary tumour and analysed their VAFs at remission.

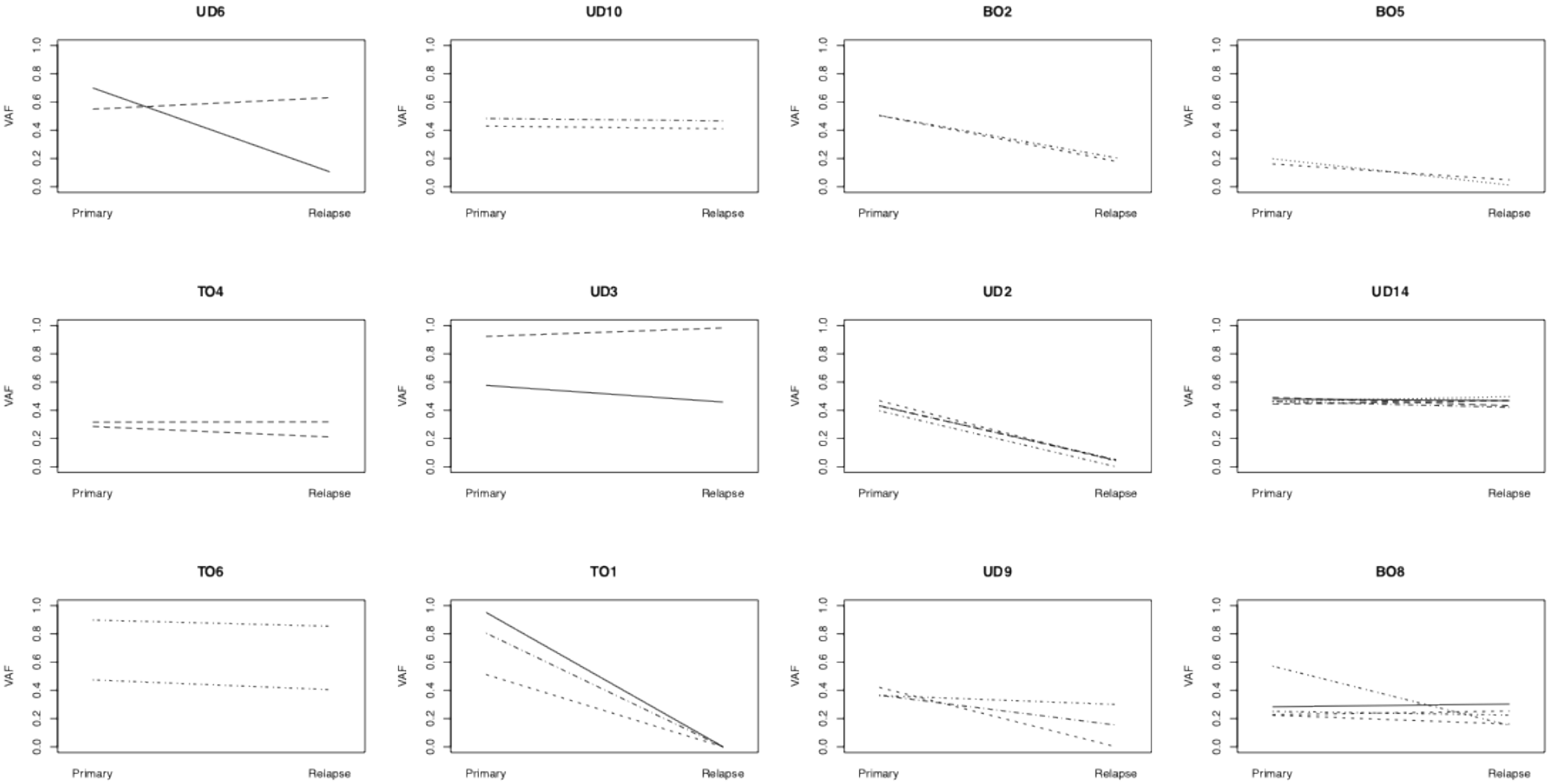


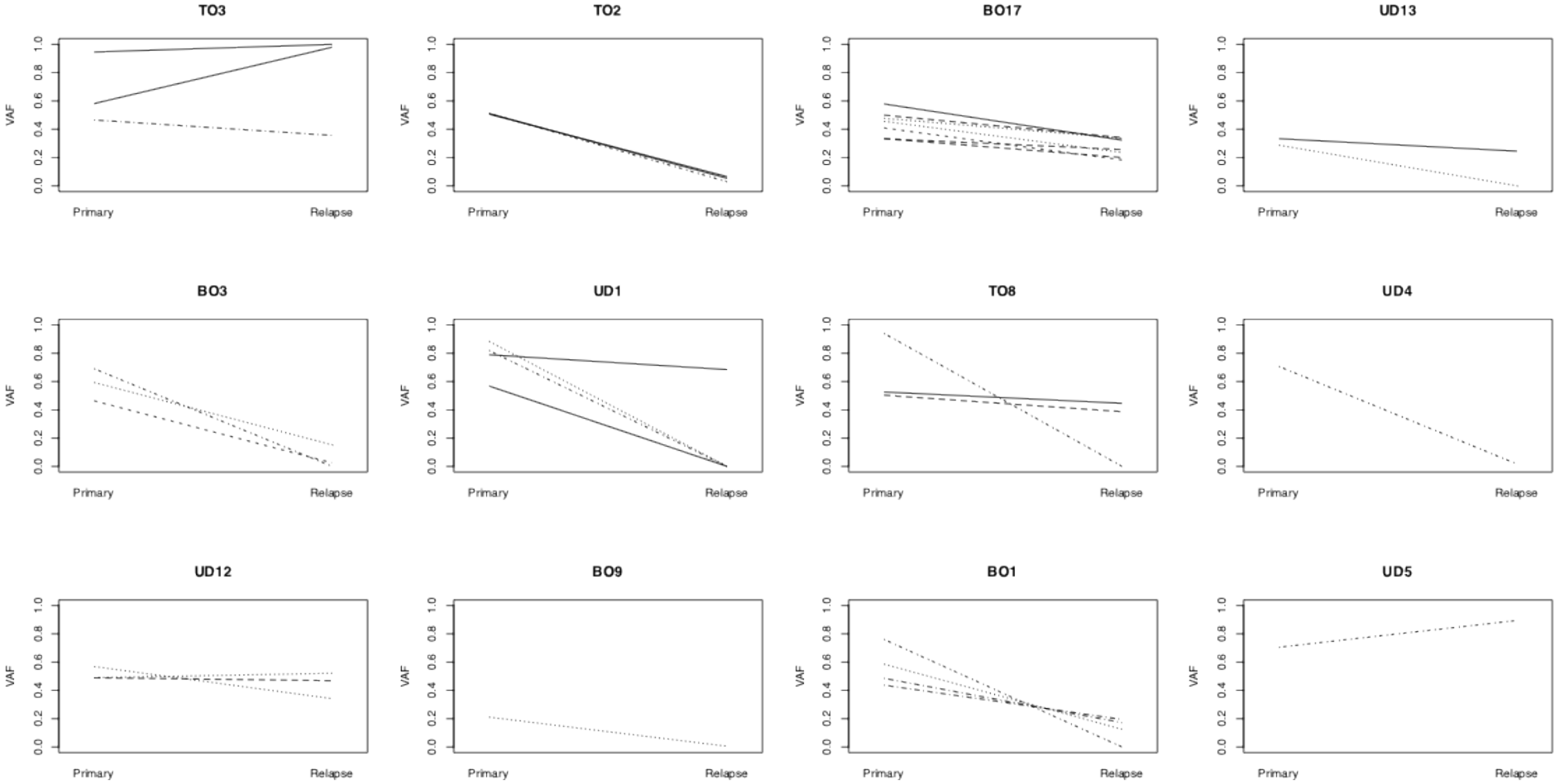
4.2.5.5 Founder mutations can be depleted at relapse

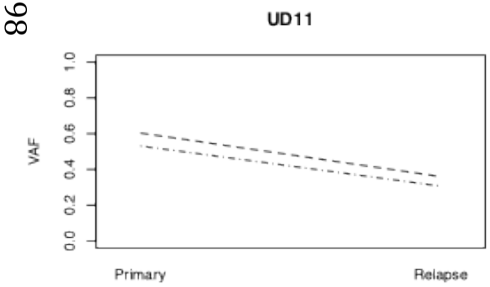
A founder clone is defined biologically as the clone harbouring the first transforming mutations, which gave rise to the clonal hierarchy forming the tumour population. If the relapse cells originated from preleukemic or primary leukaemia cells, then necessarily we would retrieve that same mutations in both primary and relapse samples at high frequency. With the aim to trace the evolution of the founder clone during the progression of the leukaemia, we

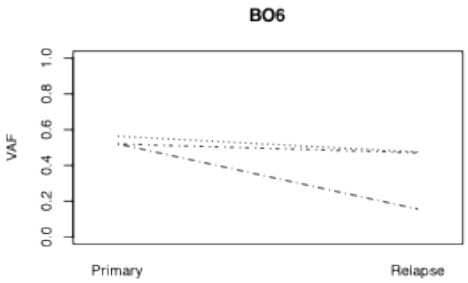
experimentally defined as founder mutations the nucleotide variants with VAFs scoring in the ninth decile of all frequencies in the primary tumour and followed their evolution trough their abundance in the cells in the different phases of the disease (Figure 4.48). Surprisingly, we observed that the founder clone is not always retained at relapse. The emblematic example is the patient TO1, in which, confirming the results obtained by the analysis of clonal composition by PyClone, we observed the complete depletion of the founder clone at relapse.

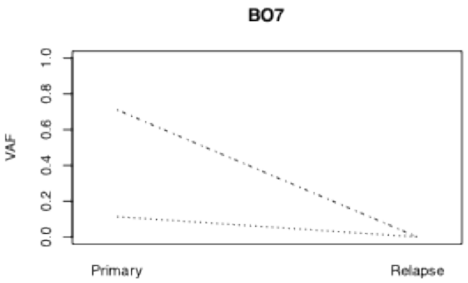
Figure 4.48: Founder mutations can be depleted at relapse. For every patient we plotted the frequencies of the founder mutations in the primary and relapse samples. In each plot, each line identifies a different founder mutation, selected as the mutations falling in the ninth decile of VAFs in the primary tumour.

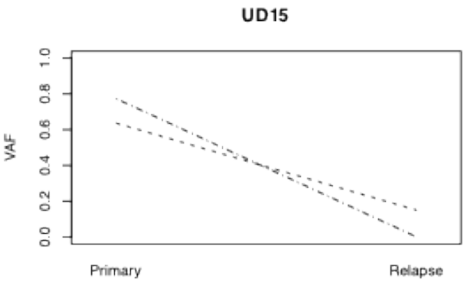

UD6
UD10
BO2
BO5
TO4
UD3
UD2
UD14
TO6
TO1
UD9
BO8
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse

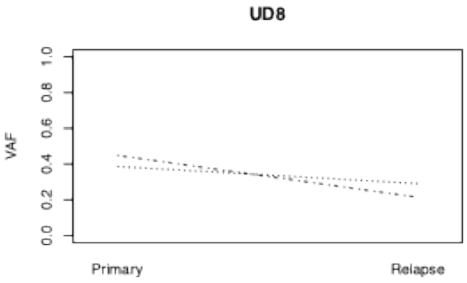

TO3
TO2
BO17
UD13
BO3
UD1
TO8
UD4
UD12
BO9
BO1
UD5
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse

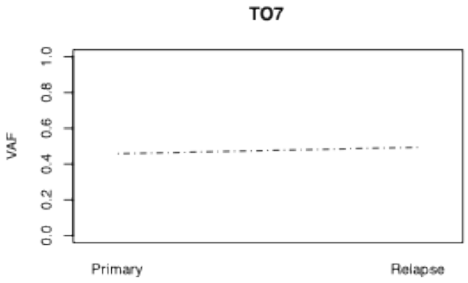

UD11
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse


BO6
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse


BO7
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse


UD15
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse


UD8
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse


TO7
VAF
0.0 0.2 0.4 0.6 0.8 1.0
Primary
Relapse

4.2.6 We were unable to validate by MiSeq relapse specific mutation in primary tumour

We next validated our SNVs using Illumina MiSeq sequencing platform: our idea was to take advantage of a deeper coverage to verify the presence of the detected mutations at lower frequencies compared to normal WES output. We therefore tested a group of variants identified in 12 patients (BO1, BO2, BO3, TO1, TO2, TO3, UD1, UD2, UD3, UD4, UD5, UD6). We choose to test 262 variants with VAFs median 40% (average 28%), belonging to the five groups defined in paragraph 4.2.3 in order to cover all the possible evolutionary characteristics. Results are reported in Table 4.12. The overall validation rate is 94% in the primary tumour (sum of "primary only", "decreasing in relapse" and "no change" groups) and 89% in the relapse (sum of "relapse only", "growing in relapse" and "no change" groups). We had the chance to test the same variants in both samples at the same time, thus we made some observations:

- Private mutations (i.e. "primary only" and "relapse only") have high validation rates: 95% and 87% in the primary and in the relapse samples, respectively;
- Common mutations (i.e. "decreasing in relapse", "no change" and "growing in relapse") also showed a high validation rate in both samples: 94% and 90% in primary and relapse samples, respectively;
- Miseq recapitulates the classes defined through WES results having a 0% of validation of "relapse only" mutations in the primary tumour and 5% of

validation of "primary only" mutations in the relapse;

- The main expectation from this analysis was, in fact, to find some relapse mutations in the primary tumour of the patients at lower frequencies. However, the MiSeq platform increases the sequencing depth but does not overcome the limits of the sequencing error introduced by the Illumina technology, preventing to distinguish very low frequency variants from errors (0.05%). It is, thus, possible that a group of "relapse only" variants is present at the exordium of the disease at frequencies lower than the detectability threshold of the MiSeq technology (0.5%) and we would need to approach this problem with a different technological solution to be able to uncover them (e.g. duplex sequencing[134]);
- the presence at low frequency of "primary only" mutations in the relapse sample, on the other hand, is the proof of the existence of a clonal heterogeneity of the tumour population, with the presence of mutations that survive chemotherapy but do not expand.

Table 4.12. Validation of 262 mutations through Illumina MiSeq sequencing platform in the primary and relapse tumour samples.

Group	Tested	Validated in primary	Validation VAF > 10%	Validation VAF < 10%	Validated in relapse	Validation VAF > 10%	Validation VAF < 10%
Primary only	65	62 (95%)	42	20	3 (5%)	1	2
Decreasing in relapse	38	37 (97%)	37	0	34 (89%)	12	22
No change	87	80 (92%)	79	1	78 (90%)	54	23
Growing in relapse	5	5 (100%)	5	0	5 (100%)	5	0
Relapse only	67	0	0	0	58 (87%)	34	24

5. Discussion

In the fight against cancer, nowadays, we have two disposable winning skills: early diagnosis and targeted drug intervention. AML is characterized by a rapid development and its early diagnosis is usually impracticable; it can be identified only rarely at beginning stages in predisposed individuals (e.g. MDS patients). The conception of novel drugs, on the other hand, can be enhanced by a better knowledge of the specific mechanisms leading to the appearance and maintenance of the pathology.

APL is the prime example of targeted therapy, advancing in about two decades from the identification of the characteristic translocation of the genes PML and RARalpha (t15;17)[135], to the first clinical trial treating APL without chemotherapeutic agents[136]. Lo Coco et al. were able to achieve 100% of remissions (on 77 patients), obtaining even better results than the association of trans-retinoic acid to idarubicin that cured patients in 95% of the cases (75/79).

For all the other AMLs, the mechanism leading to tumour development is more complex and great improvement would be accomplished by the determination of the causes of frequent therapy failures. In this thesis, endowed of the NGS technologies advancement, that revolutionized the approach for tumour investigation, we try to delineate the possible process of relapse formation with the aim, in the future, to be able to predict which patients are more susceptible to relapse.

5.1 The choice of methods for NGS analysis has to be evaluated to answer a specific question on a definite dataset

The methods for the analysis of NGS data are still in a refinement phase, at present. Indeed the scientific community is still lacking the definition of a unique pipeline combining all the best methods, especially for the high level analysis that consist in the actual detection of variants and definition of their role in the pathogenesis. In general, in order to obtain good results, it is recommended to first investigate the methods that fit best for the specific purpose of the analysis and for the specific dataset to be analysed. In fact, different methods can start from disparate assumptions that sometimes can collide with the hypothesis of the study. Moreover, the characteristics of the dataset (e.g. mutation rate, genomic instability, WGS versus WES) need to be taken in consideration when choosing the methods to analyse them.

For our project we broadly analysed the existing methods for the treatment of NGS data in order to determine those better fitting to our cohort of patients and purposes. In particular, we tested the performances of aligners, mutation callers, CNV callers and methods to reconstruct clonal composition of tumors. Occasionally, we had the possibility to choose the best tool meeting our investigative needs, discovering that other methods were valuable as well: the choice of BWA rather than Novoalign was driven mainly by an economical (BWA is not under license) aspect because the performances of the two methods were

similarly good. These results have been confirmed by the 2015 publication of Hwang et al.[137], in which the authors compared the performances of 13 WES analysis pipelines on a gold standard dataset. Also in that case BWA and Novoalign followed by GATK downstream analysis showed similarly good performances. Indeed BWA is one of the most widely used aligners for WES reads. On the contrary, for mutation calling the scientific community is making an effort trying to fill the weaknesses of the methods combining their skills in a unique tool. For this reason the ICGC-TCGA SMC-DNA meta challenge aimed at the construction of a meta-pipeline incorporating mutation calls from multiple variant callers in order to make robust variant predictions. Teams who submitted their method to the challenge organizers, in the majority of the cases, obtained levels of prediction accuracy over 90%, confirming that the aggregation of different tool's characteristics largely improves the variant detection. Despite the great advancements of the last years in the accurate identification of somatic mutations, many papers showed us evidences of intrinsic errors related to the sample preparation and analysis that challenge this task: errors introduced by PCR, imbalanced coverage[138] and sample degradation[131] can eventually result in false positive calls and many methods are nowadays under development to overcome this issue.

Certainly, the CNV detection in WES samples is tricky due to many factors: GC content and repetitive regions are probably the two more known actors but also imbalance produced by the PCR process or exome capture and the fact that some sections of the genome (i.e. regulatory regions) are generally less covered,

impair CNV calling, particularly in low coverage regions[139]. In fact Hong et al. reported a very poor reproducibility in CNV calling even from two successive analyses of the same samples, underlining the need for a careful and wide evaluation of existing methods[140]. Control FREEC, selected as best performing method in our study, also in the literature resulted to be the one showing the highest sensitivity and specificity balance for WES data[141].

Concerning clonal composition reconstruction, we concluded that none of the tested methods was ultimately satisfactory. Indeed, the extrapolation of clonal composition of the tumour population from WES data is complicated by multiple factors:

- the presence of very low frequency mutations: it is difficult to associate low frequency mutations univocally to one clone, because they can be recent mutation in any clone;
- the number of copies relative to the specific position of the mutation is a fundamental information to assess the real frequency of the mutation and the determination of the combination of frequency and number of absolute copies is not always straightforward;
- the variant frequencies are affected by the error given by the sampling of the DNA material from the whole tumour: selecting only a subset of DNA molecules from the original sample can result in overestimation or underestimation of variant frequency, especially for low frequency variants.

The ICGC-TCGA-DREAM challenge provided the opportunity to improve clonal analysis decomposition tools though the preliminary results told us that some

improvements could be made. Therefore, we made the best possible choice among the tested tools, aware that the results had to be taken with a pinch of salt.

5.2 The impact of relapse prediction in AML patients

Aiming at the characterization of the genomic landscapes of relapsing AML patients, we sequenced the primary and relapse tumour of 30 patients, and compared them with the corresponding remission sample highlighting the mutations that were retained and eventually grew after the first remission. This would allow to understand the forces that interplay in the relapse formation and to unveil the scenario that best fits with the clinical observations. AML is able to rapidly evolve and adapt to the new environmental state, thanks to the plasticity given by a genetic alteration makeup that confers an advantage to win the "struggle for life". There are many hypothetical evolutionary scenarios and it is possible that they could be overlapping for subgroups of patients. The first intriguing question is whether mutations characterizing leukemic clones are already present at diagnosis. The markers described in Paragraph 1.7 are not exhaustive for prediction of relapse formation and finding new genetic markers that can improve outcome assessment would have a great impact on the clinic. This would consequently make room for new therapeutic strategies and contemporary give to the patients the possibility to make an aware choice of the

treatment, reducing the costs for the National Health Systems and improving patients' quality of life.

Strictly related to the first question, we would like also to understand whether the relapse originated from the dominant clone, a sub-clone or possibly from a pre-leukemic clone that eventually evolved acquiring new mutations.

We observed that in the majority of the patients (76%) some relapse clones were already present in the primary tumour and reappeared at similar or augmented cellular frequencies at relapse. 21% of the patients carried clones that persisted at relapse at decreased frequency, suggesting that the therapy was not effective but the dominant clones at relapse did not originate from the more represented in the primary tumour. In a patient-wise scenario the mutations of primary and relapse tumours were different and the transition/transversion rate indicated two different mechanisms at the origin of mutations in the two groups. However at the gene level, mutation specific for the primary or relapse tumour in one patient could be common in other patients as well. The genomic landscapes of primary and relapse groups were similar and it was difficult to grasp a specific resistant genetic make-up from our cohort. Grouping genes by functional categories, we observed that genes belonging to DNA methylation pathway, cohesin complex and chromatin modifiers are prone to confer resistance to therapy.

Also CNVs patient-wise were seldom retained and, excluding those with high false positive rates, the patients with many common CNVs had also common mutations (UD4, TO4, UD1 and TO8). Recurrent mutated AML drivers were few and CNV losses were generally retained at relapse.

These results partially recapitulate previous observations of persistent "landscaping"[61] mutations at relapse; nevertheless we have noticed that cell types carrying driver mutations seldom survive chemotherapy and there is no functional category always killed by the treatment. Furthermore we remarked the fact that the founder mutations of the primary disease in some cases disappeared at relapse, posing a question mark on the origin of the clones that expanded at relapse. Indeed, at relapse we never observed clones that were completely unrelated with the primary leukaemia but the disappearance of the most frequent mutations was not compatible with the hypothesis of a hierarchical structure of tumour cells. The possible explanation for this occurrence can be either technical or biological: it is possible that the sampling made by the biopsy and sequencing does not catch the high complexity of the real situation or that the frequencies that we measure are not accurate enough to permit a reconstruction of evolution of the disease in time; from the biological point of view we can suppose that in some patients the mutations are not the founder events and can be preceded by other transforming events like CNVs, as we confirmed for two out of five patients lacking founder mutations in the relapse, or epigenetic changes.

5.3 NGS technology provides new chances for a better remission assessment

The characterization of resistant mutations in this context was tricky, probably because the cells surviving after treatment were not proliferating in the primary

tumour but owned a high proliferative potential. The typical traits of dormant cells are difficult to be defined. Indeed, we observed the emergence at relapse of clones already presented in the primary tumour, with no additional mutations at relapse; previous studies also reported this behaviour, suggesting that the dormancy phenotype of LSCs can lie outside the genetic make up of a cell.

With the situation manifested, this study has a clinical repercussion on the molecular characterization of the remission sample rather than the primary tumour. In fact, in many patients we observed that some molecular lesions peculiar of relapse sample were detectable at remission, although they were not identified in the primary tumour. Because the remission samples consisted in blood cells, we hypothesize that a NGS analysis targeted on AML driver genes would help in driving the treatment to obtain better outcomes for patients. In 4 out of 29 patients (14%) we were able to identify driver mutations at remission and in 22 out of 29 patients (76%) we detected "relapse only" mutations at low frequency in the remission sample; in these cases a targeted sequencing of the remission sample (preferentially performed through a technique allowing calling of variants at very low VAFs) would advice clinicians on the ineffectiveness of the therapy, thus allowing them to provide additional treatments when possible. A hypothetical NGS chip would contain all AML driver genes, other known drivers (CGC) and a refined list of "relapse only" mutations that we detected at remission.

5.4 Conclusion and future perspectives

The circumstances that make chemotherapy ineffective for AML treatment are complex and probably interplay on various levels: mutations, genomic rearrangements, epigenomics and regulation of dormancy/proliferation. In this thesis we identified some functional categories more prone to resistance and particular genes (not notorious in AML) presenting growing VAFs at relapse. We noticed the presence of leukemic mutations in the remission sample at low frequency and advocated the introduction of more sophisticated diagnostic tools, based on NGS technologies, to guide clinicians in the treatment decision plan.

We think that some naturally consequent steps will deepen the knowledge of relapse formation commenced in this study:

- the sequencing of primary tumour at very high sequencing depth would reveal the eventual presence of relapse mutations in the primary tumour;
- an additional cohort of AML patients that did not relapse in 5 years after the initial treatment would give better insights on the difference in the mutational landscapes of primary tumour that are more or less prone to relapse in the next five years (this cohort can be retrieved from the TCGA patients);
- the model used to build the benchmark datasets to test clonal analysis composition tools can be improved to ask a biological question: in order to test for the possibility that the presence of a pool of quiescent and somatically heterogeneous cells would favour the relapse formation, we

can simply add to the model the characteristic "dormancy", coupling it with diminished proliferation.

In the long term all the results reported in this thesis will possibly support the development and progress of new therapeutic strategies for AML.

Acknowledgements

I have to thank many people who kindly contributed to the realization of this thesis. Starting from the scientific point of view I want to thank Jole Costanza for carrying on this project and for her paramount work in the reanalysis of old samples with the new pipeline, the manual curation of mutations and indels lists, indel discovery and Figure 4.32. Also Chiara Ronchini was fundamental taking care of the sample collection and pre processing phases, giving careful advises, being always disposable to get her hands dirty and revising this manuscript. Giorgio Enrico Maria Melloni practically constructed the list of AML driver genes used throughout this thesis and helped in the complete reanalysis of TCGA-AML patients but his role has been fundamental also in the choice of statistical tests and for interesting scientific discussions. Stefano de Pretis gifted me with his great knowledge of mathematical models in order to build the benchmark for clonal analysis composition methods evaluation. Luciano Giacò Python scripted the pipeline and was a kind and precise collaborator. Anna Russo helped me practically with the complete reanalysis of TCGA-AML patients and theoretically with deep, long and late-night scientific discussions. Sara Volorio, Domenico Sardella and Loris Bernard validated with Ion Torrent the mutations we identified. I want to thank also Luca Rotta, Salvatore Bianchi and Thelma Capra for the sample sequencing and for being always very kind answering our questions and helping us in case of problems. Syed K. Hasan, Tiziana Ottone, Serena Lavorgna, Francesco Lo-Coco, Anna Candoni, Renato Fanin, Eleonora Toffoletti, Ilaria

Iacobucci, Giovanni Martinelli, Alessandro Cignetti and Corrado Tarella provided us with the patient's samples. Last but not least I want to thank my supervisors: Laura Riva and Pier Giuseppe Pelicci for giving me the possibility to make this wonderful project, for their support through the years of this PhD thesis and for helping me improving myself.

From the personal point of view I thank openhearted all the families I had and encountered during this journey: Alessandro and Arturo for making me smile everyday, for being both incredibly beautiful and for being a wonderful family; mum, dad and Iacopo for being always on my side and for reminding me who I am; Daniela, Graziano, Elisabetta, Riccardo, Giuseppe, Giovanna and Pallino for their generosity and enthusiasm; all my big family and in particular Vanessa for taking care of Arturo when I could not; "le edoline" Vera and Eugenia for sharing a part of our apartment but more importantly for sharing a part of our lives and growing together as sisters; Francesco, Giorgio, Neethu, Alberto, Nami, Valerio, Kamal, Ganesh, Luciano, Ottavio for bringing happiness in everyday work and particularly Anna, another sister who sustained me and shared all the mountains and slopes of these long and devastating 5 years.

Bibliography

1. Flaherty KT, Infante JR, Daud A, et al. Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N. Engl. J. Med.* 2012;367(18):1694-703. doi:10.1056/NEJMoa1210093.
2. Abdullah SE, Haigentz M, Piperdi B. Dermatologic Toxicities from Monoclonal Antibodies and Tyrosine Kinase Inhibitors against EGFR: Pathophysiology and Management. *Chemother. Res. Pract.* 2012;2012:351210. doi:10.1155/2012/351210.
3. ALLISON AC. Turnovers of Erythrocytes and Plasma Proteins in Mammals. *Nature* 1960;188(4744):37-40. doi:10.1038/188037a0.
4. Erslev AJ. Production of erythrocytes. In: McGraw-Hill, ed. New York; 1983:365.
5. LAJTHA LG. Stem Cell Concepts. *Differentiation* 1979;14(1-3):23-33. doi:10.1111/j.1432-0436.1979.tb01007.x.
6. Alenzi FQ, Alenazi BQ, Ahmad SY, Salem ML, Al-Jabri AA, Wyse RKH. The haemopoietic stem cell: between apoptosis and self renewal. *Yale J. Biol. Med.* 2009;82(1):7-18.
7. Ogawa M. Differentiation and Proliferation of Hematopoietic Stem Cells. *Blood* 1993;81(11):pp2844-2853.
8. Ketley NJ, Newland AC. Haemopoietic growth factors. *Postgrad. Med. J.* 1997;73(858):215-221. doi:10.1136/pgmj.73.858.215.
9. Rad A. No Title.
10. NIH. SEER Stat Facts: Leukemia.
11. Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries in 2012. *Eur. J. Cancer* 2013;49(6):1374-1403. doi:10.1016/j.ejca.2012.12.027.
12. Vedi A, Santoro A, Dunant CF, Dick JE, Laurenti E. Molecular landscapes of human hematopoietic stem cells in health and leukemia. *Ann. N. Y. Acad. Sci.* 2016;1370(1):5-14. doi:10.1111/nyas.12981.
13. Bennett JM, Catovsky D, Daniel MT, et al. Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group. *Br. J. Haematol.* 1976;33(4):451-8.
14. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood* 2009;114(5):937-51. doi:10.1182/blood-2009-03-209262.
15. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* 2016;127(20):2391-405. doi:10.1182/blood-2016-03-643544.
16. Grimwade D, Walker H, Oliver F, et al. The importance of diagnostic cytogenetics on outcome in AML: analysis of 1,612 patients entered into the MRC AML 10 trial. The Medical Research Council Adult and Children's Leukaemia Working Parties. *Blood* 1998;92(7):2322-33. doi:10.1016/0165-4608(87)90216-0.
17. Walter MJ, Payton JE, Ries RE, et al. Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc. Natl. Acad. Sci.* 2009;106(31):12950-12955. doi:10.1073/pnas.0903091106.
18. Matthews W, Jordan CT, Wiegand GW, Pardoll D, Lemischka IR. A receptor tyrosine kinase specific to hematopoietic stem and progenitor cell-enriched populations. *Cell* 1991;65(7):1143-1152. doi:10.1016/0092-8674(91)90010-V.
19. Thiede C. Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. *Blood* 2002;99(12):4326-4335. doi:10.1182/blood.V99.12.4326.
20. Colombo E, Marine J-C, Danovi D, Falini B, Pelicci PG. Nucleophosmin regulates the stability and transcriptional activity of p53. *Nat. Cell Biol.* 2002;4(7):529-33. doi:10.1038/ncb814.

21. Falini B, Mecucci C, Tiacci E, et al. Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N. Engl. J. Med.* 2005;352(3):254-66. doi:10.1056/NEJMoa041974.
22. Falini B, Martelli MP, Bolli N, et al. Acute myeloid leukemia with mutated nucleophosmin (NPM1): is it a distinct entity? *Blood* 2010;117(4):1109-1120. doi:10.1182/blood-2010-08-299990.
23. Verhaak RGW, Goudswaard CS, van Putten W, et al. Mutations in nucleophosmin (NPM1) in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood* 2005;106(12):3747-54. doi:10.1182/blood-2005-05-2168.
24. Bos JL, Verlaan-de Vries M, van der Eb AJ, et al. Mutations in N-ras predominate in acute myeloid leukemia. *Blood* 1987;69(4):1237-41.
25. Farr CJ, Saiki RK, Erlich HA, McCormick F, Marshall CJ. Analysis of RAS gene mutations in acute myeloid leukemia by polymerase chain reaction and oligonucleotide probes. *Proc. Natl. Acad. Sci. U. S. A.* 1988;85(5):1629-33.
26. Bos JL. ras oncogenes in human cancer: a review. *Cancer Res.* 1989;49(17):4682-9.
27. Barletta E, Gorini G, Vineis P, et al. Ras gene mutations in patients with acute myeloid leukaemia and exposure to chemical agents. doi:10.1093/carcin/bgh057.
28. Neubauer A, Maharry K, Mrózek K, et al. Patients with acute myeloid leukemia and RAS mutations benefit most from postremission high-dose cytarabine: a Cancer and Leukemia Group B study. *J. Clin. Oncol.* 2008;26(28):4603-9. doi:10.1200/JCO.2007.14.0418.
29. Grossmann V, Schnittger S, Poetzinger F, et al. High incidence of RAS signalling pathway mutations in MLL-rearranged acute myeloid leukemia. *Leukemia* 2013;27(9):1933-6. doi:10.1038/leu.2013.90.
30. Pabst T, Mueller BU, Zhang P, et al. Dominant-negative mutations of CEBPA, encoding CCAAT/enhancer binding protein-alpha (C/EBPalpha), in acute myeloid leukemia. *Nat. Genet.* 2001;27(3):263-70. doi:10.1038/85820.
31. Snaddon J, Smith ML, Neat M, et al. Mutations ofCEBPA in acute myeloid leukemia FAB types M1 and M2. *Genes, Chromosom. Cancer* 2003;37(1):72-78. doi:10.1002/gcc.10185.
32. Smith ML, Cavenagh JD, Lister TA, Fitzgibbon J. Mutation of *CEBPA* in Familial Acute Myeloid Leukemia. *N. Engl. J. Med.* 2004;351(23):2403-2407. doi:10.1056/NEJMoa041331.
33. Fenaux P, Preudhomme C, Quiquandon I, et al. Mutations of the P53 gene in acute myeloid leukaemia. *Br. J. Haematol.* 1992;80(2):178-183. doi:10.1111/j.1365-2141.1992.tb08897.x.
34. Kornblau S, Andreeff M, Hu S, et al. Low and maximally phosphorylated levels of the retinoblastoma protein confer poor prognosis in newly diagnosed acute myelogenous leukemia: a prospective study. *Clin. Cancer Res.* 1998;4(8):1955-1963.
35. Sugimoto K, Hirano N, Toyoshima H, et al. Mutations of the p53 gene in myelodysplastic syndrome (MDS) and MDS-derived leukemia. *Blood* 1993;81(11):3022-6.
36. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* 2008;456(7218):66-72. doi:10.1038/nature07485.
37. Sasaki M, Knobbe CB, Munger JC, et al. IDH1(R132H) mutation increases murine haematopoietic progenitors and alters epigenetics. *Nature* 2012;488(7413):656-9. doi:10.1038/nature11323.
38. Moran-Crusio K, Reavie L, Shih A, et al. Tet2 Loss Leads to Increased Hematopoietic Stem Cell Self-Renewal and Myeloid Transformation. *Cancer Cell* 2011;20(1):11-24. doi:10.1016/j.ccr.2011.06.001.
39. Ley TJ, Ding L, Walter MJ, et al. DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* 2010;363(25):2424-2433. doi:10.1056/NEJMoa1005143.
40. Bejar R, Stevenson KE, Caughey BA, et al. Validation of a prognostic model and the impact of mutations in patients with lower-risk myelodysplastic syndromes. *J. Clin. Oncol.* 2012;30(27):3376-82. doi:10.1200/JCO.2011.40.7379.
41. Mar BG, Bullinger L, Basu E, et al. Sequencing histone-modifying enzymes identifies UTX

mutations in acute lymphoblastic leukemia. *Leukemia* 2012;26(8):1881-3. doi:10.1038/leu.2012.56.
42. Makishima H, Visconte V, Sakaguchi H, et al. Mutations in the spliceosome machinery, a novel and ubiquitous pathway in leukemogenesis. *Blood* 2012;119(14):3203-10. doi:10.1182/blood-2011-12-399774.
43. Hahn CN, Venugopal P, Scott HS, Hiwase DK. Splice factor mutations and alternative splicing as drivers of hematopoietic malignancy. *Immunol. Rev.* 2015;263(1):257-278. doi:10.1111/imr.12241.
44. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* 2013;368(22):2059-74. doi:10.1056/NEJMoa1301689.
45. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013;499(7457):214-8. doi:10.1038/nature12213.
46. Melloni GE, Ogier AG, de Pretis S, et al. DOTS-Finder: a comprehensive tool for assessing driver genes in cancer genomes. *Genome Med.* 2014;6(6):44. doi:10.1186/gm563.
47. Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 2012;22(8):1589-98. doi:10.1101/gr.134635.111.
48. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 2011;18(3):507-22. doi:10.1089/cmb.2010.0265.
49. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500(7463):415-21. doi:10.1038/nature12477.
50. Champion KM, Gilbert JGR, Asimakopoulos FA, Hinshelwood S, Green AR. Clonal haemopoiesis in normal elderly women: implications for the myeloproliferative disorders and myelodysplastic syndromes.
51. Busque L, Patel JP, Figueroa M, et al. Recurrent Somatic TET2 Mutations in Normal Elderly Individuals With Clonal Hematopoiesis. doi:10.1038/ng.2413.
52. Jacobs KB, Yeager M, Zhou W, et al. Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* 2012;44(6):651-8. doi:10.1038/ng.2270.
53. Pløen GG, Nederby L, Guldberg P, et al. Persistence of *DNMT3A* mutations at long-term remission in adult patients with AML. *Br. J. Haematol.* 2014;167(4):478-486. doi:10.1111/bjh.13062.
54. Jaiswal S, Fontanillas P, Flannick J, et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. 2014;26(25). doi:10.1056/NEJMoa1408617.
55. Genovese G, Kähler AK, Handsaker RE, et al. Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* 2014;371(26):2477-2487. doi:10.1056/NEJMoa1409405.
56. Jaiswal S, Fontanillas P, Flannick J, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* 2014;371(26):2488-98. doi:10.1056/NEJMoa1408617.
57. Anastasi J, Feng J, Le Beau MM, Larson RA, Rowley JD, Vardiman JW. Cytogenetic Clonality in Myelodysplastic Syndromes Studied With Fluorescence In Situ Hybridization: Lineage, Response to Growth Factor Therapy, and Clone Expansion.
58. Raza A, Gezer S, Mundle S, et al. Apoptosis in bone marrow biopsy samples involving stromal and hematopoietic cells in 50 patients with myelodysplastic syndromes. *Blood* 1995;86(1):268-76.
59. Raza A, Galili N. The genetic basis of phenotypic heterogeneity in myelodysplastic syndromes. *Nat. Rev. Cancer* 2012;12(12):849-59. doi:10.1038/nrc3321.
60. Jan M, Snyder TM, Corces-Zimmerman MR, et al. Clonal Evolution of Preleukemic Hematopoietic Stem Cells Precedes Human Acute Myeloid Leukemia.
61. Corces-Zimmerman MR, Hong W-J, Weissman IL, Medeiros BC, Majeti R. Preleukemic mutations in human acute myeloid leukemia affect epigenetic regulators and persist in remission. *Proc. Natl. Acad. Sci.* 2014;111(7):2548-2553. doi:10.1073/pnas.1324297111.
62. Challen GA, Sun D, Jeong M, et al. Dnmt3a is essential for hematopoietic stem cell differentiation. *Nat. Genet.* 2012;44(1):23-31. doi:10.1038/ng.1009.
63. Sykes SM, Kokkaliaris KD, Milsom MD, Levine RL, Majeti R. Clonal evolution of preleukemic

hematopoietic stem cells in acute myeloid leukemia. *Exp. Hematol.* 2015;43(12):989-992. doi:10.1016/j.exphem.2015.08.012.
64. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature* 2013;502(7471):333-9. doi:10.1038/nature12634.
65. Andor N, Graham TA, Jansen M, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* 2015;22(1):105-113. doi:10.1038/nm.3984.
66. Yates LR, Campbell PJ. Evolution of the cancer genome. *Nat. Rev. Genet.* 2012;13(11):795-806. doi:10.1038/nrg3317.
67. Schramm A, Köster J, Assenov Y, et al. Mutational dynamics between primary and relapse neuroblastomas. *Nat. Genet.* 2015;47(8):872-877. doi:10.1038/ng.3349.
68. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 1987;4(4):406-25.
69. Andor N, Harness J V, Müller S, Mewes HW, Petritsch C. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics* 2014;30(1):50-60. doi:10.1093/bioinformatics/btt622.
70. Zare H, Wang J, Hu A, et al. Inferring Clonal Composition from Multiple Sections of a Breast Cancer. Tanay A, ed. *PLoS Comput. Biol.* 2014;10(7):e1003703. doi:10.1371/journal.pcbi.1003703.
71. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* 2014;11(4):396-8. doi:10.1038/nmeth.2883.
72. Miller CA, White BS, Dees ND, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. Beerenwinkel N, ed. *PLoS Comput. Biol.* 2014;10(8):e1003665. doi:10.1371/journal.pcbi.1003665.
73. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;472(7341):90-94. doi:10.1038/nature09807.
74. Greenman CD, Pleasance ED, Newman S, et al. Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.* 2012;22(2):346-361. doi:10.1101/gr.118414.110.
75. Nik-Zainal S, Van Loo P, Wedge DC, et al. The life history of 21 breast cancers. *Cell* 2012;149(5):994-1007. doi:10.1016/j.cell.2012.04.023.
76. Nik-Zainal S, Loo P Van, Wedge DC, et al. The Life History of 21 Breast Cancers. 2012;11(10):16-1. doi:10.1016/j.cell.2012.04.023.
77. Dohner K, Paschka P. Intermediate-risk acute myeloid leukemia therapy: current and future. *Hematology* 2014;2014(1):34-43. doi:10.1182/asheducation-2014.1.34.
78. Mayer R, Davis R, Schiffer C, Berg D. Intensive postremission chemotherapy in adults with acute myeloid leukemia. *Engl. J. …* 1994.
79. Döhner H, Estey EH, Amadori S, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* 2010;115(3):453-74. doi:10.1182/blood-2009-07-235358.
80. Horowitz MM, Gale RP, Sondel PM, et al. Graft-Versus-Leukemia Reactions After Bone Marrow Transplantation.
81. Estey E. Acute Myeloid Leukemia and Myelodysplastic Syndromes in Older Patients. *J. Clin. Oncol.* 2007;25(14):1908-1915. doi:10.1200/JCO.2006.10.2731.
82. Rao A V., Valk PJM, Metzeler KH, et al. Age-Specific Differences in Oncogenic Pathway Dysregulation in Patients With Acute Myeloid Leukemia. *J. Clin. Oncol.* 2009;27(33):5580-5586. doi:10.1200/JCO.2009.22.2547.
83. de Greef GE, van Putten WLJ, Boogaerts M, et al. Criteria for defining a complete remission in acute myeloid leukaemia revisited. An analysis of patients treated in HOVON-SAKK co-operative group studies. *Br. J. Haematol.* 2005;128(2):184-191. doi:10.1111/j.1365-2141.2004.05285.x.
84. de Lima M, Strom SS, Keating M, et al. Implications of potential cure in acute myelogenous leukemia: development of subsequent cancer and return to work. *Blood* 1997;90(12):4719-24. doi:10.1016/0145-2126(91)90124-c.
85. Yanada M, Garcia-Manero G, Borthakur G, Ravandi F, Kantarjian H, Estey E. Potential cure of acute myeloid leukemia. *Cancer* 2007;110(12):2756-2760. doi:10.1002/cncr.23112.

86. Swirsky DM, Bastos M de, Parish SE, Rees JKH, Hayhoe FGJ. Features affecting outcome during remission induction of acute myeloid leukaemia in 619 adult patients. *Br. J. Haematol.* 1986;64(3):435-453. doi:10.1111/j.1365-2141.1986.tb02200.x.
87. Kottaridis PD, Gale RE, Frew ME, et al. The presence of a FLT3 internal tandem duplication in patients with acute myeloid leukemia (AML) adds important prognostic information to cytogenetic risk group and response to the first cycle of chemotherapy.
88. Gale RE, Green C, Allen C, et al. The impact of FLT3 internal tandem duplication mutant level, number, size, and interaction with NPM1 mutations in a large cohort of young adult patients with acute myeloid leukemia. *Blood* 2008;111(5):2776-84. doi:10.1182/blood-2007-08-109090.
89. Thol F, Schlenk RF, Heuser M, Ganser A. How I treat refractory and early relapsed acute myeloid leukemia. *Blood* 2015;126(3).
90. Landau DA, Carter SL, Getz G, Wu CJ. Clonal evolution in hematological malignancies and therapeutic implications. *Leukemia* 2014;28(1):34-43. doi:10.1038/leu.2013.248.
91. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012;481(7382):506-10. doi:10.1038/nature10738.
92. Ding L, Getz G, Wheeler DA, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 2008;455(7216):1069-75. doi:10.1038/nature07423.
93. Szikriszt B, Póti Á, Pipek O, et al. A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* 2016;17(1):99. doi:10.1186/s13059-016-0963-7.
94. Krönke J, Bullinger L, Teleanu V, et al. Clonal evolution in relapsed NPM1-mutated acute myeloid leukemia. *Blood* 2013;122(1):100-8. doi:10.1182/blood-2013-01-479188.
95. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012;481(7382):506-510. doi:10.1038/nature10738.
96. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324.
97. Novocraft. Available at: http://www.novocraft.com.
98. Burrows M, Wheeler D I G I T A L DJ. A Block-sorting Lossless Data Compression Algorithm. 1994.
99. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 1970;48(3):443-453. doi:10.1016/0022-2836(70)90057-4.
100. Ruffalo M, LaFramboise T, Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics* 2011;27(20):2790-6. doi:10.1093/bioinformatics/btr477.
101. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20(9):1297-303. doi:10.1101/gr.107524.110.
102. Broad Institute. Picard tools.
103. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 2013;31(3):213-9. doi:10.1038/nbt.2514.
104. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38(16):e164. doi:10.1093/nar/gkq603.
105. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 2012;28(3):311-7. doi:10.1093/bioinformatics/btr665.
106. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* 2011;39(17):e118. doi:10.1093/nar/gkr407.
107. Broad Institute. Best Practices - mapping. Available at:

https://www.broadinstitute.org/gatk/guide/bp_step.php?p=1.
108. Klambauer G, Schwarzbauer K, Mayr A, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 2012;40(9):e69. doi:10.1093/nar/gks003.
109. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012;28(10):1307-13. doi:10.1093/bioinformatics/bts146.
110. Sathirapongsasuti JF, Lee H, Horst BAJ, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;27(19):2648-54. doi:10.1093/bioinformatics/btr462.
111. Boeva V, Popova T, Bleakley K, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 2012;28(3):423-5. doi:10.1093/bioinformatics/btr670.
112. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22(3):568-76. doi:10.1101/gr.129684.111.
113. Bengtsson H, Irizarry R, Carvalho B, Speed TP. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* 2008;24(6):759-67. doi:10.1093/bioinformatics/btn016.
114. BioDiscovery. Nexus Copy Number. Available at: http://www.biodiscovery.com/nexus-copy-number/.
115. Gillesple DT, Gillespie DT. Exact Stochastic Simulation of Coupled Chemical Reactions.
116. Pineda-Krch M. GillespieSSA : Implementing the Stochastic Simulation Algorithm in *R. J. Stat. Softw.* 2008;25(12):1-18. doi:10.18637/jss.v025.i12.
117. Roth A, Khattra J, Yap D, et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* 2014;11(4):396-398. doi:10.1038/nmeth.2883.
118. Miller CA, White BS, Dees ND, et al. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. Beerenwinkel N, ed. *PLoS Comput. Biol.* 2014;10(8):e1003665. doi:10.1371/journal.pcbi.1003665.
119. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* 2012;40(21):e169-e169. doi:10.1093/nar/gks743.
120. Davoli T, Xu AW, Mengwasser KE, et al. Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell* 2013;155(4):948-962. doi:10.1016/j.cell.2013.10.011.
121. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. *Science* 2013;339(6127):1546-58. doi:10.1126/science.1235122.
122. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25(14):1754-60. doi:10.1093/bioinformatics/btp324.
123. De Grassi A, Iannelli F, Cereda M, et al. Deep sequencing of the X chromosome reveals the proliferation history of colorectal adenomas. *Genome Biol.* 2014;15(8):437. doi:10.1186/s13059-014-0437-8.
124. Ruffalo M. SEAL: SEquence ALignment evaluation suite.
125. Wang Q, Jia P, Li F, et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.* 2013;5(10):91. doi:10.1186/gm495.
126. ICGC-TCGA. ICGC-TCGA DREAM Genomic Mutation Calling Challenge. Available at: https://www.synapse.org/#!Synapse:syn312572/wiki/58893.
127. Govindan R, Ding L, Griffith M, et al. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* 2012;150(6):1121-34. doi:10.1016/j.cell.2012.08.024.
128. Bodini M, Ronchini C, Giacò L, et al. The hidden genomic landscape of acute myeloid leukemia: subclonal structure revealed by undetected mutations. *Blood* 2015;125(4):600-5. doi:10.1182/blood-2014-05-576157.
129. Futreal PA, Coin L, Marshall M, et al. A census of human cancer genes. *Nat. Rev. Cancer* 2004;4(3):177-83. doi:10.1038/nrc1299.
130. Wellcome Trust Sanger Institute. Cancer Gene Census.
131. Costello M, Pugh TJ, Fennell TJ, et al. Discovery and characterization of artifactual

mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* 2013;41(6):e67-e67. doi:10.1093/nar/gks1443.
132. Chen L, Liu P, Evans TC, Ettwiller L. DNA damage is a major cause of sequencing errors, directly confounding variant identification. doi:10.1101/070334.
133. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* 2012;481(7382):506-10. doi:10.1038/nature10738.
134. Kennedy SR, Schmitt MW, Fox EJ, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* 2014;9(11):2586-2606. doi:10.1038/nprot.2014.170.
135. Alcalay M, Zangrilli D, Fagioli M, et al. Expression pattern of the RAR alpha-PML fusion gene in acute promyelocytic leukemia. *Proc. Natl. Acad. Sci. U. S. A.* 1992;89(11):4840-4.
136. Lo-Coco F, Avvisati G, Vignetti M, et al. Retinoic acid and arsenic trioxide for acute promyelocytic leukemia. *N. Engl. J. Med.* 2013;369(2):111-21. doi:10.1056/NEJMoa1300874.
137. Hwang S, Kim E, Lee I, et al. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* 2015;5:17875. doi:10.1038/srep17875.
138. Alioto TS, Buchhalter I, Derdak S, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* 2015;6:10001. doi:10.1038/ncomms10001.
139. Zhao M, Wang Q, Wang Q, et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 2013;14(Suppl 11):S1. doi:10.1186/1471-2105-14-S11-S1.
140. Hong CS, Singh LN, Mullikin JC, et al. Assessing the reproducibility of exome copy number variations predictions. *Genome Med.* 2016;8(1):82. doi:10.1186/s13073-016-0336-6.
141. Alkodsi A, Louhimo R, Hautaniemi S. Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief. Bioinform.* 2015;16(2):242-54. doi:10.1093/bib/bbu004.